

UNCLASSIFIED

AD

418387

DEFENSE DOCUMENTATION CENTER

FOR

SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION, ALEXANDRIA, VIRGINIA



UNCLASSIFIED

NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

AFCRL-63-98

64-3

CATALOGED BY DDC
AS 4691

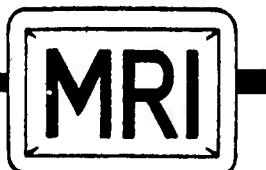
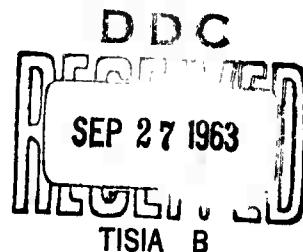
418387

FINAL REPORT
PATTERN RECOGNITION
AND
DETECTION BY MACHINES
by

Arthur E. Laemmel

Final Report No. PIBMRI-1130-63
for

Air Force Cambridge Research Laboratories
Office of Aerospace Research
United States Air Force
Bedford, Massachusetts
Contract No. AF-19(628)-375
Project 4691, Task 469103
8 March 1963



POLYTECHNIC INSTITUTE OF BROOKLYN
MICROWAVE RESEARCH INSTITUTE

ELECTROPHYSICS DEPARTMENT

"Requests for additional copies by Agencies of the Department of Defense, their contractors, and other Government agencies should be directed to the:

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

"All other persons and organizations should apply to the:

U. S. DEPARTMENT OF COMMERCE
OFFICE OF TECHNICAL SERVICES
WASHINGTON 25, D. C."

AFCRL-63-98

Research Report No. PIBMRI-1130-63

FINAL REPORT

"Pattern Recognition and Detection by Machine"

by

Arthur E. Laemmel

Polytechnic Institute of Brooklyn
Microwave Research Institute
55 Johnson Street
Brooklyn 1, New York

Final Report No. PIBMRI-1130-63

Contract No. AF-19(628)-375

8 March 1963

Project 4691

Task 469103

Title Page
Acknowledgment
Abstract
Table of Contents
62 Pages of Text
Distribution List

Arthur E. Laemmel
Arthur E. Laemmel
Research Associate Professor

Approved by:

H. J. Carlin
H. J. Carlin
Head, Electrophysics Department

Prepared for

Air Force Cambridge Research Laboratories
Office of Aerospace Research
United States Air Force
Bedford, Massachusetts

ACKNOWLEDGMENT

The work reported herein was sponsored by the Air Force
Cambridge Research Laboratories, Office of Aerospace Research
under Contract No. AF-19(628)-375.

ABSTRACT

Pattern recognition is considered generally, but emphasis is placed on the following points: optimum estimation of statistical parameters so as to minimize the probability of incorrect classification, non-Gaussian and non-stationary situations, pattern detection in a continuing time series and calculation of error probabilities. Some of the work is specifically directed toward the problem of radio station recognition. A design procedure for pattern recognizing machines is suggested which uses results from this report and other referenced sources.

Table of Contents

| | <u>Pages</u> |
|--|--------------|
| 1.1 Introduction | 1 |
| 2.1 Pattern recognition - general | 2 |
| 2.2 Basic model for pattern recognition | 3 |
| 2.3 Uncorrelated Gaussian model | 9 |
| 2.4 Model using "ramp" densities | 11 |
| 2.5 Non-stationary model | 14 |
| a) By time dependent densities | 14 |
| b) By a-priori knowledge from previous trials | 15 |
| 2.6 Modifications of the basic model | 17 |
| a) Simplification due to independent properties | 17 |
| b) Generalization: some parameters common to all classes | 17 |
| c) Generalization: Dependent property vectors | 19 |
| 2.7 Controls and regression | 20 |
| 3.1 Probability of misclassification | 23 |
| 3.2 Communication theory | 27 |
| 3.3 Added error due to finite calibration sample | 28 |
| 3.4 Game theory aspects | 31 |
| 4.1 Pattern detection | 35 |
| 4.2 Matched filters | 35 |
| 5.1 Radio station recognition | 44 |
| 5.2 Radio station recognition results | 45 |
| 5.3 Apparatus | 48 |
| 6.1 General design procedure for pattern recognizers | 53 |
| Appendix | |
| References | |

1. Introduction

The objective of the present contract is to study the applicability of machine learning processes to military detection and recognition problems. Specifically, it is desired to investigate the concept and design of a pattern recognizing machine which accepts incomplete data subject to measurement errors, which maintains an up-to-date catalog of properties of previously identified objects, and which displays the recommended classification as to the identity of unknown objects together with the probability that the decision is correct. The machine should also determine if a new pattern fits any of the categories already observed, or if it belongs to a new category previously unobserved. These objectives have been paraphrased from the "Statement of Work".

Most of the work which has been done on this contract has centered on a specific example of pattern recognition, the identification of radio stations from their carrier fading curves. This example was chosen because it allows samples to be taken quite easily in the laboratory, and because it is closely related to certain practical problems of interest to the U. S. Air Force. The present contract continued the work of a previous contract, AF-19(604)-6154, which is reported in References 13, 17, 18, 19 and 20. The present final report is supplemented by 3 scientific reports, References 21, 22 and 23.

2.1. Pattern recognition-general

It has been convenient for most workers in this field to divide a pattern recognizing machine into two parts; the first part extracts certain numerical properties from the pattern, the second part makes the decision as to which class the pattern belongs when it is given the set of properties. The present section and most of this report is concerned with the second of these two parts. The basic process can be described in non-mathematical terms, at least in its simpler aspects. Suppose that four samples of patterns are available from each of two classes, A and B. If two properties of each pattern are measured, the samples might fall as shown in Fig. 1.

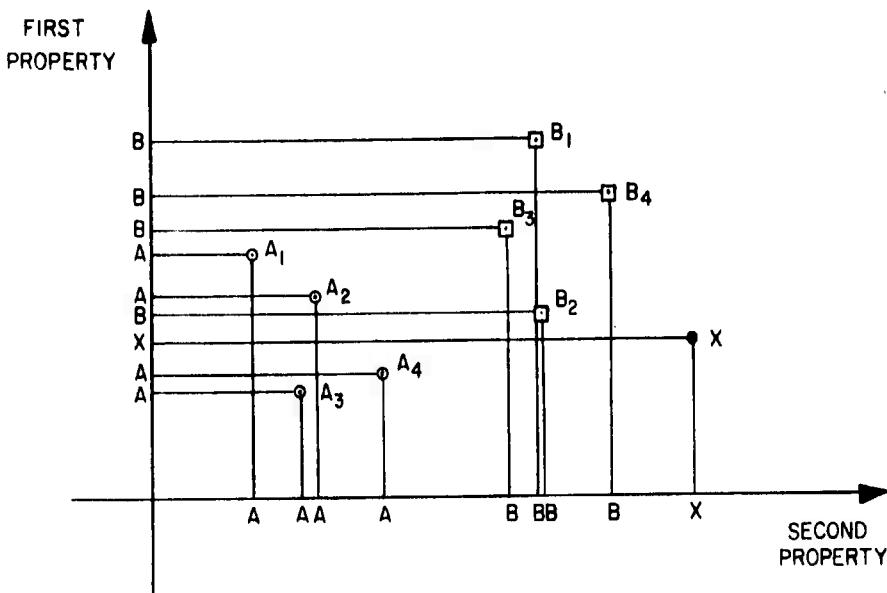


Fig.1

Now if a property vector X is observed to fall as shown, most people would not hesitate to say that it belongs to class B. The diagram also shows that the second property alone is sufficient to identify X as belonging to class B, but that the first property used alone would probably misclassify X as belonging to class A. The projection of the points on the first property axis shows an overlap of class A and B samples, whereas the projection on the second property axis shows two distinct clusters.

Some general ideas can be obtained from the above example: If the sample points fall in definite clusters which are well separated from each other, and if the unknown falls near one of these clusters, the classification is easy. If such clustering does not take place, the addition of more properties might then cause it. It is very difficult to visualize such processes in a space of many dimensions; and intuition is not always sufficient to design an optimum machine, especially where the clusters

overlap slightly. The following sections are devoted to a mathematical theory of this phase of pattern recognition which is derived from classical statistics and from communication and radar engineering. Before going on to statistical methods, it should be mentioned that some methods for pattern recognition are essentially non-statistical^{9, 27, 28}, but these can usually be considered within the statistical framework as special cases.

2.2. Basic model for pattern recognition

A mathematical framework which is suited to many types of decision making in property space will first be developed generally, then specialized applications will be made and the connections with other theories pointed out. The pattern recognizers considered here will be designed from the outset to minimize the probability of misclassification. Although it may become necessary to compromise the optimum design for various practical reasons, the relation of the usual methods to the ideal will thus be more clearly seen.

The basic difficulty in pattern recognition stems from the fact that the patterns are generated by a stochastic process, not a determinate process. That is, every time the pattern generator sets out to form a pattern of a certain class the result differs slightly from all previous attempts. This fact is described in the mathematical theory by making the property vector \underline{x} a random variable governed by the probability density $p(\underline{x}|\underline{a}_j)$, where \underline{a}_j is a vector of statistical parameters (means, variances, covariances, etc.) pertaining to class j . In almost all pattern recognition situations the true values of the \underline{a}_j are unknown, and in fact even the functional form of $p(\underline{x}|\underline{a}_j)$ is unknown. In order to have a tractable mathematical model, it will be assumed here that the form of $p(\underline{x}|\underline{a}_j)$ is known. This assumption is not a significant restriction on the generality of the theory since, if enough parameters are used, $p(\underline{x}|\underline{a}_j)$ can be made to approximate an arbitrary function to within given tolerances. For example: If $p(\underline{x}|\underline{a}_j)$ is assumed to be a Gaussian density and if \underline{x} has r components, then \underline{a} will have $\frac{r(r+3)}{2}$ components and the best choice of these may not match the actual density of \underline{x} . However, if the actual density is found to be non-Gaussian then the number of parameters can be increased to include not only means and variances but also skewness, excess etc.⁽⁵⁾ In fact, the components of \underline{a}_j might even be made to be semi-invariants, coefficients in a Gram-Charlier or Edgeworth expansion⁽⁵⁾, or in fact coefficients in any series of orthogonal functions. As will be seen below, there is a definite disadvantage in most practical applications of making the $p(\underline{x}|\underline{a}_j)$ functional form too general, i.e., of allowing too many components in the vector \underline{a}_j .

Most pattern recognition schemes require an estimation of the parameters \underline{a}_j from the given samples of the property vectors. In the mathematical model proposed here, the \underline{a}_j are themselves considered to be random variables and their values need not be estimated explicitly. A choice of the values of \underline{a}_j is assumed to be made by nature before any calibration or pattern recognition takes place, and these values are assumed to then remain fixed throughout any single experiment*. The stochastic process which generates the \underline{a}_j will be described by the density $A(\underline{a}_1, \underline{a}_2 \dots \underline{a}_k)$ where there are K classes, and this density can be used to introduce any a-priori knowledge the pattern recognizer may have of the \underline{a}_j before the calibration phase begins. Such a-priori knowledge may exist because of a physical analysis of the pattern generation mechanism, from previous attempts at pattern recognition in a time-varying environment, or for any other reason. As a special case, $A(\underline{a})$ can be considered as a flat density over a very large interval of \underline{a}_j indicating no a-priori knowledge. Since a priori knowledge of the values of the \underline{a}_j is usually insufficient, additional knowledge must be supplied through the measurement of a certain number of samples of the property vector whose class is given. The samples of class j will be denoted by $\underline{\xi}_j$, the double underline indicating that several values of each component of the property vector are given, i.e., $\underline{\xi}$ is a function of three indices, property, sample number, and class. If the sample vector selections are independent for all classes and times, the joint density for $\underline{\xi}_j$ is

$$\prod_{t=1}^d \prod_{i=1}^r p(\xi_{ji}^{(t)} | \underline{a}_j)$$

The usual maximum likelihood estimates of the \underline{a}_j would be those values which maximize the above expression. Since there is no assurance these values minimize the misclassification probability, they will not be used.

A complete pattern recognition experiment will consist of four stochastic selections (some of which are multiple): selection of the statistical parameters by "nature", selection of the calibration samples of the property vectors, selection of the class of the unknown in the actual recognition phase, and selection of the property vector. The first and third of these can be looked on as causes of the second and fourth, which might then be called effects. This situation can then be analyzed in terms of Bayes' rule as in Appendix 1. The effects are known to the pattern recognizer, the causes are unknown. If the a-priori probability of class j is p_j

*This condition is qualified somewhat in Section 2.5 below.

then the joint probability of causes a_j and effects ξ , x is

$$P(c, e) = p_j A(a_1, a_2 \dots a_k) \prod_{t=1}^d \prod_{i=1}^k p(\xi_i^{(t)} | a_i) p(x | a_j) \quad (1)$$

Since causes a_j are not really desired, the above expression should be integrated over all values of a_j . The joint probability density of desired cause and observed effects is then

$$P(j, e) = p_j \int \dots \int A(a_1, \dots, a_k) \prod_{t=1}^d \prod_{i=1}^k p(\xi_i^{(t)} | a_i) d a_1 \dots d a_k \quad (2)$$

This can be simplified if the a_j are assumed to be statistically independent:

$$P(j, e) = p_j \int p(x | a_j) \prod_{t=1}^d p(\xi_j^{(t)} | a_j) A_j(a_j) d a_j \prod_{i \neq j}^k \int \prod_{t=1}^d p(\xi_i^{(t)} | a_i) A_i(a_i) d a_i \quad (3)$$

The product with the $i = j$ term missing can be put in more convenient form by using

$$\prod_{i \neq j}^k b_i = \frac{\prod_{i=1}^k b_i}{b_j}$$

the result being

$$P(j, e) = p_j \left\{ \frac{\int p(x | a) \prod_{t=1}^d p(\xi_j^{(t)} | a) A_j(a) d a}{\int \prod_{t=1}^d p(\xi_j^{(t)} | a) A_j(a) d a} \right\} G(\xi) \quad (4)$$

where

$$G(\xi) = \prod_{i=1}^k \int \prod_{t=1}^d p(\xi_i^{(t)} | a) A_i(a) d a$$

Note that $G(\xi)$ is the joint density of all calibration data. If P is divided by the joint density of calibration data and property vector (of unknown pattern), then the conditional probability of the cause of the unknown given the calibration data and unknown property vector will be obtained. The calibration data (ξ) and unknown property vector (x) are statistically independent if the parameters a are known, but in the present model they are not. The joint density of ξ and x is given by integrating Eq. (1) over all values of $a_1 \dots a_k$ and summing for $j = 1, 2 \dots k$; this is the same as summing Eq. (2) for $j = 1, 2 \dots k$; but this is the same as summing Eq. (4). Let the ratio in Eq. (4) be denoted by $\mu(x|j, \xi)$, and the joint probability of ξ and x by $q(\xi, x)$, then the conditional probability of the unknown class being j given the calibration data ξ and the unknown property vector x is

$$q(j|x, \xi) = \frac{p_j \mu(x|j, \xi) G(\xi)}{q(x, \xi)} \quad (5)$$

where $q(x, \xi) = \sum_{h=1}^k p_h \mu(x|h, \xi) G(\xi)$

The factor G drops out, giving

$$q(j|x, \xi) = \frac{p_j \mu(x|j, \xi)}{q(x|\xi)} \quad (6)$$

where $q(x|\xi) = \sum_{h=1}^k p_h \mu(x|h, \xi)$

and $\mu(x|j, \xi) = \text{ratio of Eq. (4)}$

Reference to Appendix 1 (with $j \ a \ x \ \xi$ equivalent to $c_1 c_2 e_1 e_2$ respectively) shows that that the following are the interpretations of the functions appearing in Eq. (6): since $q(x|\xi)$ is the conditional probability density of the unknown property vector x given the calibration data ξ , therefore $\mu(x|j, \xi)$ is the conditional probability density of the unknown property vector x given its class, and given the calibration data ξ .

The connection between the pattern recognizer which results from the above model and the more usual forms may be seen quite easily. The identification of class, j , made by the above model is that value of j which maximizes Eq. (4), or if a factor not dependent on j is dropped, which maximizes:

$$\Lambda_j(\underline{x}; \underline{\xi}) = p_j \frac{\int p(\underline{x}|\underline{a}) \prod_{t=1}^d p(\xi_j^{(t)}|\underline{a}) A_j(\underline{a}) d\underline{a}}{\int \prod_{t=1}^d p(\xi_j^{(t)}|\underline{a}) A_j(\underline{a}) d\underline{a}} \quad (8)$$

The usual maximum likelihood estimate of \underline{a}_j which maximizes

$$L(\xi_j^{(1)}, \xi_j^{(2)}, \dots, \xi_j^{(d)}; \underline{a}_j) = \prod_{t=1}^d p(\xi_j^{(t)}|\underline{a}_j) \quad (9)$$

See Cramér⁽⁵⁾, Chaps. 32, 33 and especially Section 33.2. Denote the maximum likelihood estimate of \underline{a} for class j by \underline{a}_j^* . If the sample size is large, i.e., if $d \gg 1$, then L is a sharply peaked function of \underline{a} , that is L as a function of \underline{a} will approximate a Dirac δ function, $\delta(\underline{a} - \underline{a}_j^*)$. Under these circumstances, since $\int \delta(\underline{a} - \underline{a}^*) f(\underline{a}) d\underline{a} = f(\underline{a}^*)$, Eq. (8) reduces to

$$\Lambda_j(\underline{x}; \underline{\xi}) = p_j \frac{p(\underline{x}|\underline{a}_j^*) A_j(\underline{a}_j^*)}{A_j(\underline{a}_j)} = p_j p(\underline{x}|\underline{a}_j^*) \quad (10)$$

If \underline{a}_j^* happens to be the true value of \underline{a}_j , then the value of j which maximizes this expression represents the best possible choice, but if \underline{a}_j^* is a maximum likelihood estimate from a finite sample there is no reason to believe that maximizing Λ_j will be equivalent to maximizing Λ_j (Equation 8). However, it was shown above that maximizing Eq. (8) maximizes the probability of a correct choice of unknown class j , and it is shown in Appendix 1 maximizing Eq. (8) will minimize the probability of misclassification in a series of trials of the pattern recognizer. On the latter point; see also Section 3.2 below. It has been stated in at least one place in the literature⁽⁶⁾ that a maximum likelihood estimate of \underline{a} will result in a maximum likelihood estimate of j ; but this is not so since, although j is a single valued function of \underline{a} , \underline{a} is certainly not a single valued function of j . In fact, even the coordinates of the class regions in \underline{x} space do not uniquely determine the set of \underline{a} 's. A more common statement seems to be that it is convenient to first get a maximum likelihood estimate of \underline{a}_j^* and then to use them in Eq. 10 (see for example Rao⁽⁷⁾, p. 289, or Anderson⁽⁸⁾, p. 137). Note that if the true values of \underline{a}_j are known to be \underline{a}_j^0 , then $A_j(\underline{a}_j) = \delta(\underline{a}_j - \underline{a}_j^0)$ and Eq. (8) becomes $p_j p(\underline{x}|\underline{a}_j^0)$.

A more symmetrical form for Eq. (8) can be obtained as follows:

$$\Lambda_j(\underline{x}; \underline{\xi}) = p_j \frac{\int p(\underline{x}|\underline{a}) Z_j(\underline{a}) A_j(\underline{a}) d\underline{a}}{\int Z_j(\underline{a}) A_j(\underline{a}) d\underline{a}} \quad (11)$$

$$\text{where } Z_j(\underline{a}) = \prod_{t=1}^d p(\xi_j^{(t)} | \underline{a})$$

Here A_j represents the knowledge of \underline{a} before the calibration, Z_j represents the additional knowledge gained about \underline{a} during calibration, and the whole formula represents the best way to combine the a-priori knowledge and the a-posteriori knowledge in making the final recognition. If the calibration sample is very small A_j will play a more important role.

Some people seem to object to considering statistical parameters such as the \underline{a} 's as random variables. If the objection is philosophical, then it can be answered by enlarging the physical process to include parameter selection. Many statistical statements are made about runs of heads in coin-tossing where the probability of a head is p , but p itself could be a random variable if the coin to be tossed is first selected by random choice from a box of coins with various biases. If the objection is on the more practical grounds that the a priori probability density of the parameters $A(\underline{a})$ is not known in practice, then it can be answered that in many cases it is known (from previous experiments, physical analysis, etc.), but in any event one can always set $A(\underline{a})$ equal to a constant over a wide enough range of \underline{a} to cause it to drop out of Eq. (8) if no initial knowledge of the parameters is available. A Leica can always be set to behave like a box camera if desired. It might be pointed out that the method of pattern recognition leading to Eq. (8) differs from the usual one in two ways, averaging over all parameters values instead of using one estimated value, and including an a-priori probability density for the parameters. Note that the second feature could be obtained without the first (as well as the other way around) by modifying the classical method⁽⁵⁾ of parameter estimation to maximize

$$A(\underline{a}) \prod_{t=1}^d p(\xi_j^{(t)} | \underline{a})$$

The argument over whether to include $A(\underline{a})$ or not here is essentially the same the fruitless arguments over the validity of Bayes' rule. Several sources of semantic confusion also exist; the term random "variable" does not imply something which varies, the parameter \underline{a} has an expected value if it is a random variable, and hence the "mean of the mean" exists, etc.

2.3. Uncorrelated Gaussian model

In order to provide a definite example of the model given above, consider the case where the property vector has r components which are uncorrelated and normally distributed:

$$p(\underline{x} | m, \sigma) = \frac{1}{(2\pi)^{\frac{r}{2}} \prod_{i=1}^r \sigma_{ji}} e^{-\frac{1}{2} \sum_{i=1}^r \frac{(x_i - m_{ji})^2}{\sigma_{ji}^2}} \quad (12)$$

Here there are $2r$ statistical parameters (a) consisting of means (m) and standard deviations (σ) related as follows for class j :

$$a_{j1} = m_{j1} \quad a_j, r+1 = \sigma_{j1}$$

$$a_{j2} = m_{j2} \quad a_j, r+2 = \sigma_{j2}$$

$j = 1, 2, \dots, k$

$$\dots \quad \dots \quad \dots \quad \dots$$

$$a_{jr} = m_{jr} \quad a_j, 2r = \sigma_{jr}$$

The functions $A_j(a)$ will be taken as constants over the region of interest, i.e., no a-priori information about the parameters is assumed. The right side of Eq. 8 can now be written as a product of similar ratios, one for each property:

$$\Lambda_j = p_j \prod_{i=1}^r \frac{\int_{-\infty}^{\infty} \int_0^{\infty} p(x_i | m, \sigma) \prod_{t=1}^d p(\xi_{ji}^{(t)} | m, \sigma) dm d\sigma}{\int_{-\infty}^{\infty} \int_0^{\infty} \prod_{t=1}^d p(\xi_{ji}^{(t)} | m, \sigma) dm d\sigma} \quad (13)$$

where $p(x | m, \sigma)$ is a single variable Gaussian density function.

The integrations are easier if the substitution $\beta = 1/\sigma$ is made:

$$\Lambda_j = \frac{p_j}{\sqrt{2\pi}} \prod_{i=1}^r \frac{\int_{-\infty}^{\infty} \int_0^{\infty} \beta^{d-1} e^{-\frac{\beta^2}{2} \left[(d+1)m^2 - 2m(L_{ji} + x_i) + S_{ji} + x_i^2 \right]} dm d\beta}{\int_{-\infty}^{\infty} \int_0^{\infty} \beta^{d-2} e^{-\frac{\beta^2}{2} \left[dm^2 - 2m L_{ji} + S_{ji} \right]} dm d\beta}$$

$$\text{where } L_{ji} = \sum_{t=1}^d \xi_{ji}^{(t)} \text{ and } S_{ji} = \sum_{t=1}^d \xi_{ji}^{(t)2} \quad (14)$$

The integration with respect to m can be carried out by completing the square and noting that the result is simply the normal density over all values, i.e., unity. The integration with respect to β can then be carried out by noting that

$$\int_0^\infty \beta^b e^{-g\beta^2} d\beta = \frac{\Gamma\left(\frac{b+1}{2}\right)}{2g \frac{b+1}{2}} \quad \text{for } b \geq 0$$

where Γ is the Gamma function.

The result is:

$$\Lambda_j = \frac{p_j}{\sqrt{2\pi}} \sqrt{\frac{d}{d+1}} \frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d-2}{2}\right)} \prod_{i=1}^r \frac{\left[S_{ji} - \frac{L_{ji}^2}{d}\right]^{\frac{d-2}{2}}}{\left[\left(S_{ji} + x_i^2\right) \cdot \frac{(L_{ji} + x_i)^2}{d+1}\right]^{\frac{d-1}{2}}} \quad (15)$$

for $d > 2$.

This result is more easily visualized if expressed in terms of the sample mean M and sample standard deviation D defined as follows:

$$M_{ji} = \frac{1}{d} L_{ji} \quad D_{ji}^2 = \frac{1}{d} \left(S_{ji} - \frac{1}{d} L_{ji}^2 \right) \quad (16)$$

The unknown belongs to the class which maximizes

$$P_j \prod_{i=1}^r \frac{1}{D_{ji}} \left[\frac{1}{1 + \frac{1}{d+1} \left(\frac{x_i - M_{ji}}{D_{ji}} \right)^2} \right]^{\frac{d-1}{2}} \quad (17)$$

A factor depending only on d has been dropped since it cannot influence the choice of class j . Note that no approximations have been made beyond assuming normal uncorrelated distributions, statistical independence of various observed properties, and that no a-priori knowledge of the means and variances is available. In particular, the division by d in defining M and D will not affect the classification made but

only the form of expression 17; it does not amount to an estimate of the means and variances. For large d the bracket in expression 17 approaches the exponential limit, and expression 17 becomes

$$\frac{p_j}{\prod_{i=1}^r D_{ji}} e^{-\frac{1}{2} \left(\frac{x_i - M_{ji}}{D_{ji}} \right)^2}$$

which is just what would be obtained if the sample means and variances were inserted in the property vector probability density. This shows that the usual method is asymptotically optimum in minimizing classification errors, since the use of expressions 8 and 17 have been shown to be optimum for any sample size under the assumptions made here.

The integrations required to get an explicit formula for the correlated Gaussian case are very difficult, but it is not yet known that a closed form for the score function is impossible. The principal difficulty is in integrating a complicated version of Eq. 14 over all values of the parameters (means, variances, and covariances) which result in a positive definite covariance matrix.

2.4 Model using "ramp" densities. A certain aspect of the Gaussian model, either as described above or in the more usual estimator - recognizers, is quite unsatisfying intuitively. Suppose there are two equally probable classes and one property, and that the second class has a larger mean than the first. The set of x 's which will be identified as class 1 will be the set satisfying

$$D_1 \left[1 + \frac{1}{d+1} \left(\frac{x - M_1}{D_1} \right)^2 \right]^{\frac{d-1}{2}} < D_2 \left[1 + \frac{1}{d+1} \left(\frac{x - M_2}{D_2} \right)^2 \right]^{\frac{d-1}{2}}$$

If $d = 2$, this simplifies to

$$D_1^2 + \frac{(x - M_1)^2}{3} < D_2^2 + \frac{(x - M_2)^2}{3} \quad (18)$$

$$x < \frac{M_1 + M_2}{2} + \frac{3}{2} \frac{D_2^2 - D_1^2}{M_2^2 - M_1^2}$$

The first term places the boundary point halfway between the two centers of gravity of the sample points; the second term is a correction for the case where the variances of the two classes are different. The difficulty is that the correction is always

12.

toward the class with the larger variance. To illustrate why this is disturbing, consider

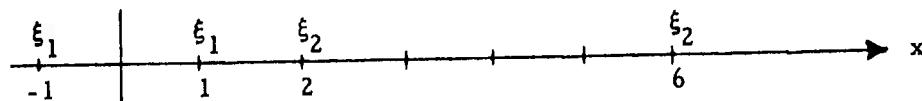
$$\begin{array}{lll} \xi_1^{(1)} = -1 & \xi_1^{(2)} = 1 & \xi_2^{(1)} = 2 \\ & & \xi_2^{(2)} = 5 \\ M_1 = \frac{-1 + 1}{2} = 0 & & M_2 = \frac{2 + 5}{2} = 3.5 \\ D_1^2 = \frac{1 + 1}{2} = 1 & & D_2^2 = \frac{2.25 + 2.25}{2} = 2.25 \end{array}$$

and the region identified as class 1 is

$$x < \frac{0 + 3.5}{2} + \frac{3}{2} \frac{2.25 - 1}{2}$$

$$x < 2.93$$

Thus, the recognizer will call one of the samples of the second class, $\xi_2^{(1)}$, a member of the first class! This does not seem reasonable, for if the points are plotted



almost anyone who is asked would place the boundary between +1 and +2; eq. 1.9 should, it seems, be called a member of class 2. The basic difficulty is in assuming a Gaussian distribution. If the properties are not normally distributed the above process is not known to be optimum, and two points is certainly much too small a sample to give the slightest verification of whether the distribution is indeed Gaussian.

Some workers⁽⁹⁾ have essentially abandoned the statistical approach, not only for this reason but for others as well. But it seems desirable for many reasons to attempt to retain the statistical framework. Before developing this approach it should be emphasized that the boundary found in the example above ($x = 2.93$) is quite reasonable if it is certain that the distributions are Gaussian, since the maximum number of points on the x axis may be classified correctly if one of the sample points is put on the wrong side of the boundary. In such a case the recognizer is bound to have a fairly high error rate and the only question is to keep it as low as possible.

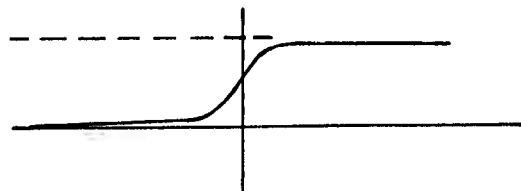
If a person is given samples of several classes of pattern vectors in two dimensions and asked to draw boundary lines between the regions which are to be classified into each of the classes, he would probably try to minimize the number of sample points on the wrong sides of boundaries,* while at the same time keeping the

*In Reference 9 some theoretical justifications for doing this are given.

boundary curves reasonably simple. Of course, by using complicated curves and separated regions all sample points could be correctly classified, but the probability of correctly classifying unknown vectors would not thereby be increased. In carrying out this process, there is a tendency to emphasize the importance of points near the boundaries and to almost ignore points far from any boundary. This is in contrast to "Gaussian" methods, which weight the importance of points according to how much they contribute to the knowledge of the means (equal weights) and to knowledge of the variances (weighted more the farther from the mean).

It is now desired to reconcile the intuitive approach illustrated above with the use of Eq. 8 as a score function. This can be done by finding a probability density $p(x|a)$ which emphasizes the sample points near the boundary of the two classes. Note that if a large sample is obtained the probability density may not turn out to resemble the one used here, but it must be repeated that the aim is to improve the pattern recognition process, not to estimate probability density functions. The ramp density function which is to be given may not improve the recognition accuracy, but it will enable a mathematical comparison between non-parametric intuitive approaches and more formal Gaussian approaches.

First, define a continuous function $\psi(x)$ with the following properties: $\psi(x)$ approaches 1 monotonically as x approaches infinity. The function $\psi(x) - \frac{1}{2}$ is an odd function of x . This implies that $\psi(0) = \frac{1}{2}$ and that $\psi(x)$ approaches zero as x approaches minus infinity. The function $\psi(x)$ looks like a step function with a finite rise time:



The function $\psi(x)$ is not integrable and hence cannot be a probability density. However, a suitable density can be defined in terms of $\psi(x)$:

$$p(x|\mu, \beta) = b \psi[\beta(x - \mu)] e^{-\epsilon|x|} \quad (19a)$$

Here ϵ is a very small number chosen so that $\epsilon|x|$ is small for any x which is likely to arise, b is a constant chosen to insure that the integral of p is unity ($-\infty < x < \infty$), μ is a parameter locating the position of the step, and β is a second statistical parameter whose magnitude gives the (inverse) rise time and whose sign gives the direction of rise (to right or left). If the above density is substituted in

Eq. 8 there results

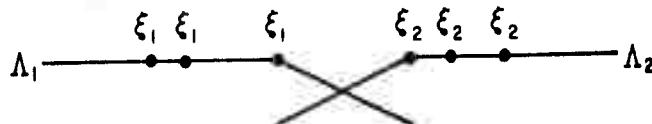
$$\Lambda_j(x) = b e^{-\epsilon|x|} \frac{\int_{-\infty}^{\infty} \psi \left[(-)^j \beta_0 (x - \mu) \right] \prod_{t=1}^d \psi \left[(-)^j \beta_0 \left(\xi_j^{(t)} - \mu \right) \right] a(\mu) d\mu}{\int_{-\infty}^{\infty} \prod_{t=1}^d \psi \left[(-)^j \beta_0 \left(\xi_j^{(t)} - \mu \right) \right] a(\mu) d\mu} \quad (19b)$$

provided the following assumption are made: that β can assume only the value $-\beta_0$ for class $j = 1$ and the value β_0 for class $j = 2$, and that $a(\mu)$ is the a-priori density of μ for both classes. Now consider the function $\Lambda_2(x)$ with $\epsilon = 0$ and with $\beta_0 \rightarrow \infty$. The integrand in the denominator vanishes unless μ is less than $\xi_2^{(1)}$, $\xi_2^{(2)} \dots \xi_2^{(d)}$ and the integrand in the numerator vanishes unless μ is less than x , $\xi_2^{(1)}, \xi_2^{(2)} \dots \xi_2^{(d)}$.

Therefore,

$$\Lambda_2(x) = \begin{cases} = b & \text{if } x > \xi_2^{(t)} \text{ for some } t \\ = \frac{\bar{a}(x)}{\bar{a}(\xi_2^{(\min)})} & \text{if } x < \xi_2^{(t)} \text{ for all } t \end{cases} \quad (19c)$$

where $\bar{a}(u) = \int_{-\infty}^u a(\mu) d\mu$. The following diagram shows that if all the $\xi_2^{(t)}$ are larger than every $\xi_1^{(t)}$, and if $a(\mu)$ is almost constant in the region of interest, the boundary will be placed between the groups:



If the sample points overlap, β_0 must be set equal to some large value in order to avoid an indeterminate boundary.

2.5 Non-stationary model a) The generalization (of the basic model of Section 2.2) which will be developed here is useful for cases where the statistics of the pattern generator vary as functions of time. Let $\tau_1, \tau_2 \dots \tau_d$ be a series of points in time at each of which one sample property vector for each class is available. The τ 's are to be arranged in increasing order, and τ is to be a time later than any of the calibration times at which the recognition is to be made. Assume that the joint

density of the properties is $p(x|\underline{a}; \tau)$ at time τ , and that the joint density of all calibration data is now

$$\prod_{t=1}^d p\left[\underline{\xi}_j^{(t)} \mid \underline{a}_j; \tau_t\right] \quad (20)$$

All of the steps of Section 2.2 can be carried thru analogously leading to the following instead of Eq. 4:

$$P(j, e) = p_j \frac{\int p(x|\underline{a}, \tau) \prod_{t=1}^d p\left(\underline{\xi}_j^{(t)} \mid \underline{a}; \tau_t\right) A_j(\underline{a}) d\underline{a}}{\int \prod_{t=1}^d p\left(\underline{\xi}_j^{(t)} \mid \underline{a}; \tau_t\right) A_j(\underline{a}) d\underline{a}} G\left(\frac{\underline{\xi}}{\underline{\xi}}\right) \quad (21)$$

where $G\left(\frac{\underline{\xi}}{\underline{\xi}}\right) = \prod_{i=1}^k \int \prod_{t=1}^d p\left(\underline{\xi}_i^{(t)} \mid \underline{a}; \tau_t\right) A_i(\underline{a}) d\underline{a}$

Since G does not depend on j , the classification is again that value of j which maximizes the first two factors on the right side of Eq. 21.

Some interesting questions are raised in examining the above process. None of the calibration times τ_t can be greater than τ ; and if the usual case of τ greater than all of the τ_t is considered, the pattern recognition process also involves an element of prediction. The location of the boundaries between the regions representing the classes in property space must be predicted at time τ . The boundaries depend on two probability densities, eg.

$$p_j p(x|\underline{a}_j^*; \tau) = p_i p(x|\underline{a}_i^*; \tau)$$

determines the boundary $x = \theta(\tau)$ between the i and j regions. Note that the parameters \underline{a} are not assumed to be functions of time, but that this does not essentially restrict the generality of the method. Thus, a possible form for the probability density is

$$p(x|\sigma, m, b; \tau) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} (x - m - bt)^2} \quad (22)$$

which can be looked on either as a normal distribution with constant variance σ and changing mean $m + bt$, or as a density involving three parameters which do not change with time σ, m, b .

b) Another method for handling a time-varying process is sometimes more appropriate than that given above. If the data is gathered in such a way that a fairly

large sample is obtained before the statistics change, but the recognition is to be carried out at a later time when only a small calibration sample is available, the previously-obtained knowledge can be entered thru the functions A . The classification at the later time is for class j which minimizes

$$p_j = \frac{\int \prod_{t=1}^d p\left(\frac{x}{\underline{a}} \mid \underline{a}\right) p\left(\frac{\xi_j(t)}{\underline{a}} \mid \underline{a}\right) A\left(\underline{a} - \underline{a}_j^*\right) d\underline{a}}{\int \prod_{t=1}^d p\left(\frac{\xi_j(t)}{\underline{a}} \mid \underline{a}\right) A\left(\underline{a} - \underline{a}_j^*\right) d\underline{a}} \quad (23)$$

where $A(\underline{a})$ is a suitable function peaked at $\underline{0}$ and where \underline{a}_j^* is the estimate of \underline{a} for class j made at the earlier time. If the early data is from a large sample and if conditions are assumed not to have changed too much then A can be made a sharply peaked function and the old data will be more heavily weighted. This method does not involve prediction, but it does accomplish somewhat the same thing. If p and A are both Gaussian, a closed form can be obtained for the score, but it is difficult to visualize for large d .

c) Still another way of handling non-stationary statistics is the use of controls as outlined in Section 2.7 below. This does not involve any predictions either, but rather it would be especially useful if the statistics change periodically, or return to previously estimated values in a non-periodic manner.

To illustrate the value of prediction, consider the case of 2 properties and 2 classes shown in Fig. 2:

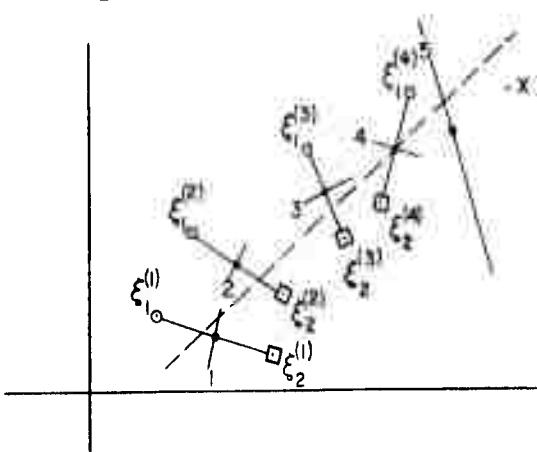


Fig. 2

If the t index in $\xi_j^{(t)}$ represents time, and if x is measured at time $t = 5$, then x should be in class 1. On the other hand, if t does not represent time, and if $\xi_1^{(1)}$ does not correspond to $\xi_2^{(1)}$, then x should be in class 2. Line 5 is the predicted or extrapolated boundary for $t = 5$.

2.6 Modifications of the basic model

The basic model for a self-calibrating pattern recognizer which was discussed in Section 2.2 above can be both specialized and generalized to suit different conditions. The recognizer is essentially described by the score function $\Lambda_j(\underline{x}, \underline{\xi})$ of Eq. 8: Given that \underline{x} is the property vector of the pattern whose class is desired, and given the calibration data $\underline{\xi}$, the best choice for the unknown's class is that value of j which maximizes Λ_j .

a) A special case which is frequently used is to assume that all of the components of the property vector are statistically independent if the parameters are known:

$$p(\underline{x} | \underline{a}) = \prod_{i=1}^r p(x_i | a_{ji})$$

where a_{ji} denotes the parameters of property i of class j . Making this substitution in Eq. 8, it is easily shown that

$$\Lambda_j(\underline{x}, \underline{\xi}) = p_j \prod_{i=1}^r \lambda_{ji}(x_i, \xi_{ji}) \quad (24)$$

$$\text{where } \lambda_{ji}(x_i, \xi_{ji}) = \frac{\int p(x_i | \underline{a}) \prod_{t=1}^d p(\xi_{ji}^{(t)} | a) A_{ji}(a) da}{\int \prod_{t=1}^d p(\xi_{ji}^{(t)} | a) A_{ji}(a) da}$$

The uncorrelated Gaussian case illustrated in Section 2.3 above was in turn a special case of this.

b) Several generalizations of Eq. 8 can be obtained, principally by relaxing some of the conditions of statistical independence. The general procedure will always be to follow Bayes' rule by writing down the joint probability density of all causes and effects and integrating out those causes which are not desired. As is shown in the Appendix, this will minimize the probability of misclassification. Unfortunately, the score function does not usually factor into as simple a form as Eq. 8. The probability density function which governs the property vector will be generalized to the form $p(\underline{x} | \underline{a}, \underline{\beta})$ where \underline{x} is a property vector of r components, \underline{a}_j is a statistical parameter vector pertinent to the j 'th class, and $\underline{\beta}$ is a vector of statistical parameters common to all classes. In order to clarify the meaning of $\underline{\beta}$, these parameters might be the variances and covariances which several workers^(8, 10) have assumed constant for all classes. The property

vectors of each of the k classes are observed d times, each such observation being a vector similar to \underline{x} :

$$\begin{array}{ccc}
 \begin{matrix} (1) \\ \underline{\xi}_1 \end{matrix} & \begin{matrix} (2) \\ \underline{\xi}_1 \end{matrix} & \begin{matrix} (d) \\ \underline{\xi}_1 \end{matrix} \\
 \begin{matrix} (1) \\ \underline{\xi}_2 \end{matrix} & \begin{matrix} (2) \\ \underline{\xi}_2 \end{matrix} & \begin{matrix} (d) \\ \underline{\xi}_2 \end{matrix} \\
 \cdots & \cdots & \cdots \\
 \begin{matrix} (1) \\ \underline{\xi}_k \end{matrix} & \begin{matrix} (2) \\ \underline{\xi}_k \end{matrix} & \begin{matrix} (d) \\ \underline{\xi}_k \end{matrix}
 \end{array}$$

The causes $j, \underline{a}_1, \underline{a}_2 \dots \underline{a}_k, \underline{\beta}$ are assumed to be statistically independent random variables, but there may be dependence between the components of the \underline{a}_i or the $\underline{\beta}$. If all of the causes are fixed, the effects $\underline{\xi}_1^{(1)}, \underline{\xi}_1^{(2)} \dots \underline{\xi}_1^{(d)}$ and \underline{x} are assumed to be independent random variables. Under these restrictions, the joint probability density of all causes and events is:

$$P(c, e) = p_j p(\underline{x} | \underline{a}_j, \underline{\beta}) B(\underline{\beta}) \prod_{i=1}^k A_i(\underline{a}_i) \prod_{t=1}^d p(\underline{\xi}_i^{(t)} | \underline{a}_i, \underline{\beta}) \quad (25)$$

The joint density of the class of the unknown j and the effects is

$$P(j, e) = p_j \int B(\underline{\beta}) \int p(\underline{x} | \underline{a}_j, \underline{\beta}) \prod_{i=1}^k A_i(\underline{a}_i) \prod_{t=1}^d p(\underline{\xi}_i^{(t)} | \underline{a}_i, \underline{\beta}) d\underline{a} d\underline{\beta} \quad (26)$$

The integrations with respect to \underline{a} can again be factored

$$P(j, e) = p_j \int B(\underline{\beta}) \frac{\int p(\underline{x} | \underline{a}, \underline{\beta}) \prod_{t=1}^d p(\underline{\xi}_j^{(t)} | \underline{a}, \underline{\beta}) A_j(\underline{a}) d\underline{a}}{\int \prod_{t=1}^d p(\underline{\xi}_j^{(t)} | \underline{a}, \underline{\beta}) A_j(\underline{a}) d\underline{a}} G(\underline{\xi}_j | \underline{\beta}) d\underline{\beta} \quad (27)$$

$$\text{where } G(\underline{\xi}_j | \underline{\beta}) = \prod_{i=1}^k \int \prod_{t=1}^d p(\underline{\xi}_i^{(t)} | \underline{a}_i, \underline{\beta}) A_i(\underline{a}) d\underline{a}$$

Note that G is the conditional density of all calibration data given the common parameters $\underline{\beta}$; and if the sample is large enough G should have a sharp peak at $\underline{\beta} = \underline{\beta}^*$, the estimated values of $\underline{\beta}$ and probably close to the true values. The function G then behaves as a delta

function and the effect is to insert $\underline{\beta}^*$ in the ratio above and in B. Then if the sample is large enough,

$$\prod_{t=1}^d p(\underline{\xi}_j^{(t)} | \underline{a}_j, \underline{\beta}^*)$$

also behaves as a delta function with a result that

$$P(j, e) \approx p_j p(\underline{x} | \underline{a}_j, \underline{\beta}^*) B(\underline{\beta}^*)$$

This shows that the recognizer described by Eq. 27 above give the same result as the ordinary method of estimating parameters in the limit of large sample size.

c) The next generalization which seems desirable is to remove the restriction that the calibration vectors $\underline{\xi}_1^{(t)}, \underline{\xi}_2^{(t)}, \dots, \underline{\xi}_k^{(t)}$ be statistically independent. In the case of speech recognition some of the properties may be related to frequencies, and a person with a low pitched voice may have the frequencies of all the phonemes lower together. In the case of radio station recognition⁽¹⁷⁾, the ionospheric conditions may be poor on a certain day (value of t) and hence the amplitudes of all stations may be low. A more general form for the joint density of causes and effects is then (replacing Eq. 1):

$$P(c, e) = p_j p(\underline{x} | \underline{a}_j) \prod_{t=1}^d P(\underline{\xi}_1^{(t)}, \underline{\xi}_2^{(t)}, \dots, \underline{\xi}_k^{(t)} | \underline{a}_1, \dots, \underline{a}_k) A(\underline{a}_1, \dots, \underline{a}_k) \quad (28)$$

Here it is still assumed that the properties are independent for different values of t. Of course, this could simply be integrated with respect to the \underline{a} and called a score function, but the resulting expression would be of little help in actual applications. There is a need to again describe each property vector by its own probability density, and this can be done by introducing an "undesired" cause or influence such as was done in the Appendix. Let γ be a random variable which describes the particular pattern generator in use at the present time; $\gamma \in \mathcal{G}$, where \mathcal{G} is the space of all pattern generators. The probability density of the property vector \underline{x} using generator γ of class j is $p(\underline{x} | \underline{a}_j, \gamma)$. The joint density of calibration vectors for a particular t is

$$P(\underline{\xi}_1^{(t)}, \dots, \underline{\xi}_k^{(t)} | \underline{a}_1, \dots, \underline{a}_k) = \int_{\mathcal{G}} \prod_{i=1}^k p(\underline{\xi}_i^{(t)} | \underline{a}_i, \gamma) \Gamma_t(\gamma) d\gamma \quad (29)$$

where $\Gamma_t(\gamma)$ is the probability density of the various generators. If Eq. 29 is inserted in Eq. 28, and if the \underline{a} 's are independent, some manipulation yields:

$$P(j, e) = p_j \int \dots \int \Gamma_1(\gamma_1) \dots \Gamma_d(\gamma_d) \frac{\int p(\underline{x}|\underline{a}) \prod_{t=1}^d p(\xi_j^{(t)}|\underline{a}; \gamma_t) A_j(\underline{a}) d\underline{a}}{\int \prod_{t=1}^d p(\xi_j^{(t)}|\underline{a}; \gamma_t) A_j(\underline{a}) d\underline{a}} G(\xi; \dots \gamma_1 \dots \gamma_d) d\gamma_1 \dots d\gamma_d \quad (30)$$

$$\text{where } G(\xi; \gamma_1 \dots \gamma_d) = \prod_{i=1}^k \int \prod_{t=1}^d p(\xi_i^{(t)}|\underline{a}; \gamma_t) A_i(\underline{a}) d\underline{a}$$

$$\text{and where } p(\underline{x}|\underline{a}) = \int p(\underline{x}|\underline{a}; \gamma) \Gamma(\gamma) d\gamma$$

Note that if γ_t is regarded as a known sampling time τ_t , then

$$\Gamma_t(\gamma_t) = \delta(\gamma_t - \tau_t) \quad (31)$$

and Eq. 30 reduces to the time-varying prediction of Eq. 21. If the functions $\Gamma_t(\gamma)$ have a sharp peak at one of two values of γ (γ_A and γ_B), depending on whether t is in one or another of two disjoint subsets of $1, 2, \dots, d$, then Eq. 30 amounts to designing two pattern recognizers, one for $\gamma = \gamma_A$ and one for $\gamma = \gamma_B$. The present method is thus useful in the usually-difficult case of multimodal probability densities, and it permits the decision region of a particular class in property space to be concave or even disconnected.

2.7 Controls and regression

In the last section a method was discussed in which it is attempted to remove the influence of an unwanted cause on the pattern recognition. Some of the properties, components of \underline{x} , might be chosen not to be influenced by the desired cause (the class of the unknown) but rather by the undesired cause. The idea is to allow these "control" properties to remove the effects of the undesired cause by means of their statistical dependence (in the Gaussian case, correlation) on the properties which depend on both the class and the undesired influence. Assume that the v 'th class is described by the following Gaussian density:

$$p(\underline{x}|\underline{a}_v) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{|\det D_{vij}|^{\frac{1}{2}}}{\prod_{i=1}^n} e^{-\frac{1}{2} \sum_{i,j=1}^n (x_i - m_{vi})(x_j - m_{vj}) D_{vij}} \quad (32)$$

where D_{vij} is the inverse of the covariance matrix, m_{vi} the means (both the D_{vij} and the

\underline{m}_v constitute the a 's), and where $|\det D_{vij}|$ is the determinant of the inverse covariance matrix. Suppose that $x_1, x_2 \dots x_\theta$ depend on the class and the undesired influence but that $x_{\theta+1}, x_{\theta+2} \dots x_n$ depend only on the undesired influence. Let the properties be designated by $y_i = x_i - m_i$ $i = 1, 2, \dots n - \theta$ and let the controls be designated by $z_i = x_{\theta+1} - m_{\theta+i}$ $i = 1, 2, \dots n - \theta$ ^{*}. If the D matrix is partitioned as

$$\begin{bmatrix} D_v \\ \vdash \end{bmatrix} = \begin{bmatrix} \begin{array}{c} \uparrow \\ \theta \\ \downarrow \\ n-\theta \end{array} & \begin{array}{c} \xleftarrow{\quad} \theta \xrightarrow{\quad} \\ S_v \\ \sim R_v \\ \hline R_v \\ \hline W \end{array} \end{bmatrix}$$

where \underline{R}_v is the transpose of $\underline{\underline{R}}_v$, it is seen that W does not depend on the class v . The quadratic form in the exponent of the density can now be expanded:

$$(\underline{\underline{x}} - \underline{\underline{m}}) \underline{\underline{D}}_v (\underline{\underline{x}} - \underline{\underline{m}}) = \underline{\underline{y}} \underline{\underline{S}}_v \underline{\underline{y}} + 2\underline{\underline{y}} \underline{\underline{R}}_v \underline{\underline{z}} + \underline{\underline{y}} \underline{\underline{W}} \underline{\underline{z}} \quad (33)$$

Now consider a change of variables:

$$\underline{\underline{y}} = \underline{\underline{y}}' + \underline{\underline{H}}_v \underline{\underline{z}}' \quad , \quad \underline{\underline{z}}' = \underline{\underline{z}}' \quad (34)$$

Since the Jacobian of this transformation is unity, the probability density of the primed variables can be obtained merely by substituting in Eq. 32. Note that

$$\begin{bmatrix} \underline{\underline{y}} \\ \underline{\underline{z}} \end{bmatrix} = \begin{bmatrix} \begin{array}{cc} \frac{1}{\underline{\underline{v}}} & \underline{\underline{H}} \\ \underline{\underline{0}} & \frac{1}{\underline{\underline{v}}} \end{array} \end{bmatrix} \begin{bmatrix} \underline{\underline{y}}' \\ \underline{\underline{z}}' \end{bmatrix}$$

and therefore the new inverse covariance matrix $\underline{\underline{D}}'$ is given by

$$\underline{\underline{D}}' = \begin{bmatrix} \begin{array}{cc} \frac{1}{\underline{\underline{v}}} & \underline{\underline{0}} \\ \sim \underline{\underline{H}} & \frac{1}{\underline{\underline{v}}} \end{array} \end{bmatrix} \begin{bmatrix} \begin{array}{cc} \underline{\underline{S}}_v & \underline{\underline{R}}_v \\ \sim \underline{\underline{R}}_v & \underline{\underline{W}} \end{array} \end{bmatrix} \begin{bmatrix} \begin{array}{cc} \frac{1}{\underline{\underline{v}}} & \underline{\underline{H}} \\ \underline{\underline{0}} & \frac{1}{\underline{\underline{v}}} \end{array} \end{bmatrix}$$

It is desired to make the $\underline{\underline{y}}'$ and the $\underline{\underline{z}}'$ statistically independent, ie. to make the off-diagonal portion of the above matrix 0. This requires that

$$\begin{aligned} \underline{\underline{S}}_v \underline{\underline{H}}_v + \underline{\underline{R}}_v \underline{\underline{z}}' &= 0 \\ \underline{\underline{H}}_v &= - \underline{\underline{S}}_v^{-1} \underline{\underline{R}}_v \end{aligned} \quad (35)$$

^{*}According to the Appendix of this report and Ref. 13, the controls should be treated just as if they were properties.

Now since

$$\underline{D}' = \begin{bmatrix} \underline{S}_{\underline{v}} & \underline{0} \\ \underline{0} & \underline{W} - \underline{\tilde{R}} \underline{S}_{\underline{v}}^{-1} \underline{R} \end{bmatrix}$$

both the determinant and the exponential term in the joint probability density of \underline{y}' and \underline{z}' will factor, and the conditional density of \underline{y}' given \underline{z}' becomes

$$\Lambda_v = \frac{|\det \underline{S}_{\underline{v}}|^{\frac{1}{2}}}{(2\pi)^{\frac{r}{2}}} e^{-\frac{1}{2} \underline{\tilde{y}}' \underline{S}_{\underline{v}}^{-1} \underline{y}'}$$

This is the proper score function for pattern recognition with controls. In terms of the original variables $x_1, x_2 \dots x_r$:

$$\Lambda_v = \frac{|\det \underline{S}_{\underline{v}}|^{\frac{1}{2}}}{(2\pi)^{\frac{r}{2}}} e^{-\frac{1}{2} (\underline{x} - \underline{m} - \underline{H}_{\underline{v}} - \underline{z})' \underline{S}_{\underline{v}}^{-1} (\underline{x} - \underline{m} - \underline{H}_{\underline{v}} \underline{z})} \quad (36)$$

where the \underline{H} are given by Eq. 35 and are the usual multiple regression coefficients⁽⁸⁾.

3.1 Probability of misclassification

All of the preceding analysis has been directed toward choosing the most probable class of a pattern, and this was done in such a way as to minimize the probability of incorrect classification. Now it is desirable to consider just how the actual error probabilities can be calculated, how large they might be, and to compare different models in this regard.

The first type of error probability will be that obtained if the true form of the probability density is known and if the parameters in this density are known. This probability of misclassification will be denoted by Q_0 . Later other error probabilities will be considered, eq. Q_1 , the probability of misclassification where the form of the probability density is known but where the values of the parameters must be estimated from a finite calibration sample.

Let there be k classes of patterns, each governed by an r variable probability density $p_i(x_1, x_2 \dots x_r)$, $i = 1, 2, \dots k$. If the density functions are known, the situation is quite simple. Let the a-priori probability of the i 'th class be p_i , then if the pattern $x_1, x_2 \dots x_r$ is observed the classification is that value of i which maximizes

$$p_i p_i(x_1, x_2 \dots x_r)$$

Except for boundary regions of zero probability, the x space can be divided into k disjoint regions $R_1, R_2 \dots R_k$ such that in region R_i the above expression is larger for class i than for any other class. The actual probability of misclassification is then

$$Q_0 = 1 - \sum_{i=1}^k p_i \iint_{R_i} \dots \int p_i(x_1, x_2 \dots x_r) dx_1 dx_2 \dots dx_r \quad (37)$$

This can be calculated directly and there is no need for an experiment. The simplest case is two normally distributed classes with $r = 1$.

$$p_i(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x - m_i)^2}{2\sigma^2}} \quad i = 1, 2$$

Taking $m_1 < m_2$ for convenience, R_1 is defined by

$$-\infty < x < \frac{m_1 + m_2}{2}$$

and R_2 by

$$\frac{m_1 + m_2}{2} < x < +\infty$$

The probability of misclassification is then given by the "error function" (defined as in Jahnke and Emde⁽¹¹⁾ and Cramer⁽⁸⁾ respectively), assuming $p_1 = p_2$:

$$Q_o = \frac{1}{2} - \frac{1}{2} \operatorname{erf}\left(\frac{m_2 - m_1}{2\sqrt{2}\sigma}\right) = 1 - \Phi\left(\frac{m_2 - m_1}{2\sigma}\right) \quad (38)$$

If $\sigma \rightarrow 0$ the error approaches zero. The worst case, $m_2 = m_1$, results in a probability of error of 1/2. Some intermediate results are tabulated below:

| $\frac{m_2 - m_1}{2\sigma}$ | 0 | .5 | .7 | 1 | 1.5 | 2 | ∞ |
|-----------------------------|----|------|------|------|------|------|----------|
| Q_o | .5 | .309 | .242 | .159 | .067 | .023 | 0 |

Table 1

Note that the density functions can be fairly large at the boundary between R_1 and R_2 . For $Q_o = .159$ the boundary is one "standard deviation" from each means, and the density function is about 60% of its maximum value. The two variable normal error function, although not appearing in every statistics book as does the one variable function, has been extensively tabulated⁽¹²⁾. The simplest case, uncorrelated variables, reduces to the two variable case with a suitable rotation of coordinates. Suppose there are r properties, each with the same separation of means $m_2 - m_1$, each with the same standard deviation σ , and all uncorrelated. Then the two classes will be separated by r times the separation in the one dimensional case, and the probability of error will be

$$Q_o = 1 - \Phi\left[\frac{\sqrt{r}(m_2 - m_1)}{2\sigma}\right] \quad (39)$$

Some representative results are tabulated below (in % this time)

| r | 0 | .5 | .7 | 1 | 1.5 | 2 | ∞ |
|-----|----|------|------|------|-----|-----|----------|
| 1 | 50 | 30.9 | 24.2 | 15.9 | 6.7 | 2.3 | 0 |
| 2 | 50 | 24.0 | 16.3 | 7.5 | 1.7 | 0.3 | 0 |
| 4 | 50 | 15.9 | 8.1 | 2.3 | 0.1 | 0 | 0 |
| 9 | 50 | 6.7 | 1.8 | 0.1 | 0 | 0 | 0 |
| 25 | 50 | 0.6 | 0 | 0 | 0 | 0 | 0 |

Table 2

The conclusion to be drawn from this case is that a moderate number of properties can be used to make an almost-perfect pattern recognizer even though each property is individually very poor, provided the different properties are uncorrelated. Since most pattern recognizers use several properties in which the means are separated by say half a standard deviation or more, and since most results are not as good as given above, the indication is that many pattern recognizers fail because the properties used are correlated with each other in an undesirable way. Of course, not all forms of correlation are deleterious; if the concentration ellipse is perpendicular to the line connecting the centroids, the performance can be improved by the correlation since a linear function of the original properties will have a very large value for $(m_2 - m_1)/\sigma$. Consider the case where there are two properties and two normally distributed classes having the following densities:

$$p_1 = \frac{1}{2\pi \sigma^2 \sqrt{1-\rho^2}} \exp \left[-\frac{x^2 - 2\rho xy + y^2}{2\sigma^2 (1-\rho^2)} \right]$$

$$p_2 = \frac{1}{2\pi \sigma^2 \sqrt{1-\rho^2}} \exp \left[-\frac{(x-m)^2 - 2\rho(x-m)(y-m) + (y-m)^2}{2\sigma^2 (1-\rho^2)} \right]$$

By integrating these with respect to x or y from $-\infty$ to $+\infty$ it will be seen that each property, when considered alone, has a variance σ^2 and a difference of means of m . If $\rho = 0$ the previous uncorrelated case results. In any case, it is seen by symmetry that the boundary between R_1 and R_2 is the 45° line $x + y = m$. The probability of error is then:

$$Q_0 = \frac{1}{2\pi \sigma^2 \sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{m-x}^{\infty} \exp \left[-\frac{x^2 - 2\rho xy + y^2}{2\sigma^2 (1-\rho^2)} \right] dx dy$$

This special case can be integrated in terms of the single variable error function:

$$Q_0 = 1 - \Phi \left[\frac{m}{\sqrt{2} \sigma \sqrt{1+\rho}} \right] \quad (40)$$

| $\rho \downarrow$ | $\frac{m}{2\sigma}$ | | | | | | |
|-------------------|---------------------|------|------|------|-----|-----|----------|
| | 0 | .5 | .7 | 1 | 1.5 | 2 | ∞ |
| 1 | 50 | 30.9 | 24.2 | 15.9 | 6.7 | 2.3 | 0 |
| 0 | 50 | 24.0 | 16.3 | 7.5 | 1.7 | 0.3 | 0 |
| -.5 | 50 | 15.9 | 8.1 | 2.3 | 0.1 | 0 | 0 |
| -1 | | 0 | 0 | 0 | 0 | 0 | 0 |

Table 3

Note that $\rho = 1$ is the worst case, but here the performance is merely degraded to that obtained for the single variable case. This fact, adding useless or correlated properties cannot degrade the performance, has been proved as a general theorem in a previous report⁽¹³⁾.

The case where there are more than two classes is very difficult to treat generally because there are so many ways to place the centroids in the r dimensional property space, each one leading to the integration of a multivariate normal density function over polytopes. One method which might be used to obtain results which can be compared with those given above is to assume that the centroids are themselves normally distributed, making the expected value of the separation of two classes a function of the standard deviation of the distribution of centroids. Some related calculations have been carried out by S. O. Rice in connection with a communication problem⁽¹⁴⁾. A simple case which can be analyzed is that of four classes with two properties, the centroids being at $(0, 0)$, $(0, m)$, $(m, 0)$ and (m, m) . If the variances are all σ^2 and the properties are not correlated, the probability of error is:

$$Q_0 = 1 - \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\frac{m}{2}} \int_{-\infty}^{\frac{m}{2}} \exp \left[-\frac{x^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} \right] dx dy \quad (41)$$

This integral can be found as a special case of the bivariate error function tabulated in Pearson⁽¹²⁾. (Actually, the above can be integrated in terms of the 1 variable error function if $\rho = 0$.) The percent of misclassified patterns is found to be:

| $\frac{m}{2\sigma}$ | 0 | .5 | .7 | 1 | 1.5 | 2 | ∞ |
|---------------------|----|------|------|------|------|-----|----------|
| Q_0 | 75 | 52.3 | 42.5 | 28.4 | 12.9 | 4.6 | 0 |

Table 4

As might be expected, the error probabilities are almost doubled over two classes in one dimension, and more than doubled over two classes in two dimensions. Suppose it is desired to recognize $k = 2^r$ classes with r properties. The probability of error is then

$$Q_o = 1 - \frac{1}{(2\pi)^{\frac{r}{2}} \sigma^r} \int_{-\infty}^{\frac{m}{2}} \cdots \int_{-\infty}^{\frac{m}{2}} \exp \left[-\frac{1}{2\sigma^2} \sum_{j=1}^r x_j^2 \right] dx_1 \cdots dx_r$$

$$Q_o = 1 - \left[\Phi \left(\frac{m}{2\sigma} \right) \right]^r \quad (41a)$$

For any fixed $m/2\sigma$, Q_o approaches unity as r approaches infinity. This is, of course, a very large number of classes to recognize; it would be very unusual for the number of classes to go up exponentially with the number of properties.

3.2 Communication theory

The theory of communication provides some useful ideas, relations and formulas for certain types of pattern recognition. Pattern recognition is similar to communication where the type of channel noise is unknown, only samples of each message are available. Note particularly the fact that random coding is often as good as the best known systematic codes, since the designer of a pattern recognizer cannot usually design the pattern generator also. Let μ^2 be the expected value of the square of one mean of a class, there being k classes in r dimensional space, and let σ^2 be the standard deviation of one of the (independent) properties. The fundamental channel capacity formula, $CT = TW \log_2 (1 + S/N)$, gives

$$\log_2 k = \frac{br}{2} \log_2 \left(1 + \frac{\mu^2}{\sigma^2} \right) \quad \text{with } b < 1 \quad (42)$$

if the signalling is to be at a rate no greater than the channel capacity. The probability of error approaches zero as $r \rightarrow \infty$ for b fixed, but the rate of approach is faster if b is smaller. The formulas for probability of error are more complicated, but if some approximations are made to Eq. 6.90 of Fano's book⁽³¹⁾ there results for small b :

$$Q_o \leq \frac{1.84 k}{\sqrt{r \log_2 \left(1 + \frac{\mu^2}{\sigma^2} \right)}} - \frac{r}{4} \log_2 \left(1 + \frac{\mu^2}{\sigma^2} \right) \quad (42a)$$

In pattern recognition the fact that $Q_o \rightarrow 0$ as $r \rightarrow \infty$ does not have the significance it does

in communication since there is seldom available an inexhaustible supply of independent properties. It would be useful to know the meanings of the analogs of Fano's Eq. 5.164 and 5.199 in pattern recognition.

3.3 Added error due to finite calibration sample

The above analyses have assumed that the statistical parameters were known quantities, but usually they are not known and must be estimated from calibration data. The probability of incorrect classification might now be expected to be larger due to the errors in parameter estimation. Suppose that the j^{th} class is governed by the probability density

$$p(x_1, x_2 \dots x_r; a_{j1}, a_{j2} \dots a_{jn})$$

The parameters a_{ji} are unknown and must be estimated by observing d property vectors from each of the k classes. Let $\xi_{j2}^{(t)}$ be the v^{th} property from the t^{th} sample from the j^{th} class. The usual maximum likelihood estimate of $a_{j1}, a_{j2} \dots a_{jn}$ is the set of values which maximizes

$$\prod_{t=1}^d p(\xi_{j1}^{(t)} \dots \xi_{jr}^{(t)}; a_{j1} \dots a_{jn}) \quad (43)$$

The estimated values of the parameters, $\underline{a}_j^*(\underline{\xi})$, are then inserted into the density and multiplied by the a-priori probabilities p_j :

$$\Lambda_j(\underline{x}, \underline{\xi}) = p_j p \left[\underline{x} \mid \underline{a}_j^*(\underline{\xi}) \right] \quad (44)$$

This may be regarded as a score function; the class with the highest score is the class of \underline{x} . Another type of score function is given by Eq. 8 above. It is often very difficult to relate the accuracy of the parameter estimation and the accuracy of the pattern recognizer, and indeed the arguments leading up to the optimum recognizer of Eq. 8 do not explicitly use parameter estimation at all. In the following analyses it is usually most convenient to take the score function as a starting point.

The probability of the pattern recognizer misclassifying a vector of class j given the calibration data $\underline{\xi}$ is

$$Q(\underline{\xi}, j) = \int p(\underline{x} \mid \underline{a}_j) d\underline{x} \quad (45)$$

$$\Lambda_j \neq \Lambda_1, \Lambda_2 \dots \Lambda_k$$

The range of integration is over that region of space where the j th score is not the largest. The expected value of this Q over all classes and calibration vectors is the desired probability of misclassification:

$$Q_1 = \sum_{j=1}^k p_j \int_{-\infty}^{\underline{\xi}} G(\underline{\xi}) \left[\int_{\underline{R}_j(\underline{\xi})} p(\underline{x} | \underline{a}_j) d\underline{x} \right] d\underline{\xi} \quad (46)$$

where $\underline{R}_j(\underline{\xi})$ is the complement of the j th class region in property space, ie. the region described in Eq. 45. Note that the $p(l)$ in Eq. 46 is the true density and need not be the same as $p(l)$ in Eq. 44. An equivalent form to Eq. 45 which may be more convenient is:

$$Q_1 = \sum_{j=1}^k p_j \int \dots \int \underline{P}_j(\Lambda_1, \Lambda_2, \dots, \Lambda_k) d\Lambda_1 \dots d\Lambda_k \quad (47)$$

$$\Lambda_j \neq \Lambda_1, \Lambda_2, \dots, \Lambda_k$$

where \underline{P}_j is the joint probability density of the scores considered as random variables on the condition the \underline{x} is chosen from the j th class. The value of Q_o , the probability of error of an optimum pattern recognizer which knows the parameter values, is never larger than Q_1 ; for:

$$\sum_{j=1}^k p_j \int_{\underline{R}_j} p(\underline{x} | \underline{a}_j) d\underline{x} \leq \sum_{j=1}^k p_j \int_{\underline{R}_j(\underline{\xi})} p(\underline{x} | \underline{a}_j) d\underline{x}$$

if the R_j are the regions which minimize the expression on the left. Multiplying by $G(\underline{\xi})$ and integrating with respect to $\underline{\xi}$:

$$Q_o \leq Q_1$$

In all of the above equations the \underline{a}_j are their true values, although these may not be known to the pattern recognizer.

A example will now be calculated to show the importance of calibration errors in a simple case. Let there be two normally distributed classes with equal variances σ^2 and described by one property. The mean of the first class will be taken as 0, the mean of the second as m . If the a-priori probabilities of the two classes are equal,

the probability of error can be calculated from one class alone:

$$Q_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_1(\Lambda_1, \Lambda_2) d\Lambda_1 d\Lambda_2$$

A possible choice for the score functions is :

$$\Lambda_1(x, \xi_1^{(1)} \dots \xi_1^{(d)}) = \frac{1}{d} (\xi_1^{(1)} + \xi_1^{(2)} + \dots + \xi_1^{(d)}) - x \quad (48)$$

$$\Lambda_2(x, \xi_2^{(1)} \dots \xi_2^{(d)}) = x - \frac{1}{d} (\xi_2^{(1)} + \xi_2^{(2)} + \dots + \xi_2^{(d)})$$

Here the decision can be made by a discriminant function $\Lambda_2 - \Lambda_1$:

$$\Delta = 2x - \frac{1}{d} (\xi_2^{(1)} + \xi_2^{(2)} + \dots + \xi_2^{(d)} + \xi_1^{(1)} + \dots + \xi_1^{(d)}) \quad (49)$$

Δ will also be normally distributed with a mean $-m$ and a variance $(4 + \frac{1}{d^2} 2d) \sigma^2$. The probability of error is now simply the probability of Δ being positive:

$$Q_1 = 1 - \Phi \left(\frac{m}{\sqrt{1 + \frac{1}{2d}} 2\sigma} \right) \quad (50)$$

The error approaches the value of Q_0 given by Eq. 38 as d approaches infinity. The increase in error % is not very significant as the following table shows.

| $\frac{m}{2\sigma}$ | 0 | .5 | .7 | 1 | 1.5 | 2 | ∞ |
|---------------------|----|------|------|------|------|-----|----------|
| d=1 | 50 | 34.2 | 28.4 | 20.7 | 11.0 | 5.1 | 0 |
| d= ∞ | 50 | 30.9 | 24.2 | 15.9 | 6.7 | 2.3 | 0 |

Table 5

An actual pattern recognizer might make more errors if the sample size is small. In the first place, if the recognizer does not know that the variances are equal, it might be tempted to apply a correction such as Eq. 18 above. Also, the scores used here assume that the mean of class 2 is greater than class 1; the decision is always for class 1 if x is below the average of the two means even if the second sample mean happens to fall below the first. The scores of Eq. 48 would only be used if it were known beforehand

that the second mean is above the first, but the actual error probabilities are not changed significantly in any case.

3.4 Game theory aspects

In attempting to design a pattern recognizing machine for property vector probability density functions more general than Gaussian, some difficulties arise even in defining just what the problem is from the mathematical viewpoint. Game theory provides a satisfactory framework in which to consider such problems. One very practical question which needs to be answered is : How many statistical parameters should be included in the assumed form of the property vector density ? If too few parameters are included the actual density cannot be approximated closely enough, if too many parameters are included their estimates made from a small calibration sample will be inaccurate or even impossible. This problem will occur in some other guise even if the pattern recognizer does not depend on explicit parameter estimation, as in Section 2.2 above.

A pattern recognizer $\mathcal{R}(d)$ will be defined as a set of rules whereby the class j of a property vector \underline{x} is decided upon after observing d sample vectors $\underline{\xi}_i^{(t)}$ of each of the k possible classes. The set of rules comprising $\mathcal{R}(d)$ may or may not be stated in terms of score functions $\Delta_j(\underline{x}, \underline{\xi})$ or assumed probability density functions. A sequence of pattern recognizers $\{\mathcal{R}(1), \mathcal{R}(2), \dots\} \equiv \mathcal{R}$ occurs if the size of the calibration sample d is increased. The choice of the pattern recognizer is the move of the first player. The actual probability density function of the property vector of the j 'th class will be assumed to be $p(\underline{x} | j)$. This is independent of d , and all of the vectors \underline{x} and $\underline{\xi}_j^{(t)}$ ($t=1, 2, \dots, d$, $j=1, 2, \dots, k$) are assumed to be independent random variables. The choice of the property vector density is the move of the second player. The second player might be considered "nature", but if "nature" is thought of as indifferent and impersonal this does not represent a very satisfying picture of a competitive game. A better choice is to regard the second player as a "devil's advocate" who is trying to encourage the designer to make a good pattern recognizer by showing how poorly suggested designs might work under the most unfavorable conditions. The utility function or payoff matrix will be taken here as the probability of misclassification given by Eq. 46 or Eq. 47. In general, the payoff or probability of misclassification will be some function of the 2 "moves" and the sample size:

$$Q_1 = Q_1 [\mathcal{R}(d), p(\underline{x} | j), d] \quad (51)$$

Games of this type are called two-person zero-sum games. The first player tries to minimize Q_1 , the second player tries to maximize Q_1 . An "equilibrium pair" is defined as a recognizer \mathcal{R}^μ and a density p_ν such that

$$Q_1 [\mathcal{R}_\mu, p]$$

attains a maximum at $p = p_\nu$, p in some fixed class, and

$$Q_1 [\mathcal{R}, p_\nu]$$

attains a minimum at $\mathcal{R} = \mathcal{R}_\mu$, \mathcal{R} in a fixed class. There may or may not be an equilibrium pair depending on the function Q_1 and on the classes of \mathcal{R} 's and p 's which are allowed. If several equilibrium pairs exist, they must have the same value for Q_1 ^(15,16). A pattern recognizer \mathcal{R}_μ which is one member of an equilibrium pair would be an optimum design for the class of probability densities considered. If no equilibrium pair exists in the sense defined above, it has been shown by von Neumann^(15,16) that the expected value of Q_1 can be minimized by adopting a "mixed strategy", ie. by choosing among the \mathcal{R} 's according to some set of probabilities instead of in a determinate manner. This achieves the same result as an equilibrium pair.

An example will now be given to illustrate the aspects of game theory which are involved here. Consider two classes with one property. The mean of the first class will be taken as 0 and the mean of the second as m . Only the case $d = \infty$ will be considered. The pattern recognizer will be described by the threshold θ : all $x < \theta$ will be called class 1, all $x > \theta$ will be called class 2. The probability densities will be either Gaussian, or flat:

$$p(x|m, \sigma_2) = \begin{cases} \frac{1}{2\sqrt{3}\sigma_2}, & -\sqrt{3}\sigma_2 < x - m < \sqrt{3}\sigma_2 \\ 0, & \sqrt{3}\sigma_2 < |x - m| \end{cases}$$

The probability of misclassification, assuming equal a-priori probabilities, is

$$Q_1 = \frac{1}{2} \left[\frac{1}{\sqrt{2\pi}\sigma_1} \int_0^\infty e^{-\frac{x^2}{2\sigma_1^2}} dx + \frac{1}{\sqrt{2\pi}\sigma_2} \int_{m-\theta}^\infty e^{-\frac{x^2}{2\sigma_2^2}} dx \right]$$

$$Q_1 = 1 - \frac{1}{2} \Phi\left(\frac{\theta}{\sigma_1}\right) - \frac{1}{2} \Phi\left(\frac{m-\theta}{\sigma_2}\right)$$

Four probability densities will be considered:

$$P_1: \begin{cases} \sigma_1 = .33 m \\ \sigma_2 = .67 m \end{cases}$$

$$P_2: \begin{cases} \sigma_1 = .1 m \\ \sigma_2 = .4 m \end{cases} \quad (\text{Gaussian})$$

$$\text{and } p_3 \begin{cases} \sigma_1 = .33 \text{ m} \\ \sigma_2 = .33 \text{ m} \end{cases} \quad p_4 \begin{cases} \sigma_1 = .25 \text{ m} \\ \sigma_2 = .25 \text{ m} \end{cases} \quad (\text{Flat})$$

Three pattern recognizers will be considered:

$$\mathcal{R}_1: \theta = \frac{m}{2} \quad \mathcal{R}_2: \theta = \frac{\sigma_1 m}{\sigma_1 + \sigma_2} \quad \mathcal{R}_3: \theta = .4 \text{ m}$$

The payoff matrix is:

| | p_1 | p_2 | p_3 | p_4 |
|-----------------|-------|-------|-------|-------|
| \mathcal{R}_1 | 14.6 | 5.3 | 6.7 | 0 |
| \mathcal{R}_2 | 15.9 | 2.3 | 6.7 | 0 |
| \mathcal{R}_3 | 15.0 | 3.3 | 7.7 | 1.9 |

Table 6

There is one equilibrium pair, (\mathcal{R}_1, p_1) .

According to game theory, \mathcal{R}_1 is preferred to the others, provided that the only probability densities considered are those shown.

In any practical situation far more probability densities must be considered than in the above example. The first question is whether equilibrium pairs exist. Q_1 can always be made at least as large as $1 - \min p_j$ for any given \mathcal{R} by the proper choice of $p(\underline{x} | j)$. To show this, set

$$p(\underline{x} | j) \begin{cases} 1/I_j(\xi) & \underline{x} \in R_h(\xi) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } I_j(\xi) = \int_{R_h(\xi)} dx$$

$$\text{and } p_h \leq p_i \quad i = 1, 2, \dots, k$$

Then, according to Eq. 46, $Q_1 = 1 - p_h$. A similar argument shows that Q_i can always be made zero by proper choice of $p(\underline{x} | j)$ for a fixed \mathcal{R} : set

$$p(\underline{x} | j) = \begin{cases} 1/I_j(\xi) & \underline{x} \in R_j(\xi) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } I_j(\underline{\xi}) = \int_{R_j(\underline{\xi})} d\underline{x}$$

and the bracketed term in Eq. 46 will vanish. Now what is the range of Q_1 with $p(\underline{x}|j)$ fixed and \mathcal{R} variable? The minimum value of Q_1 is attained by defining the R_j as follows

$$\underline{x} \in R_j^0(\underline{\xi}) \quad \text{if } p_j p(\underline{x} | j) > p_i p(\underline{x} | i) \text{ for all } i$$

$$\underline{x} \in \bar{R}_j^0(\underline{\xi}) \quad \text{if } p_j p(\underline{x} | j) \leq p_i p(\underline{x} | i) \text{ for some } i$$

Since this definition is independent of $\underline{\xi}$, Eq. 4c reduces to

$$Q_{\infty} = p(\underline{x}|j) = \sum_{j=1}^k p_j \int_{\bar{R}_j^0(\underline{\xi})} p(\underline{x} | j) d\underline{x} \quad (52)$$

The quantity Q_{∞} is, by definition, independent of \mathcal{R} and d and it may even vanish if the p 's are disjoint. The largest value of Q_1 is obtained by putting \underline{x} in R_j where j minimizes $p_i p(\underline{x}|i)$. An equilibrium pair will only exist if the row maximum is the same as the column minimum:

$$Q_{\infty} = 1 - \min_i p_i \quad (53)$$

where Q_{∞} is given by Eq. 52. This pattern recognizer would make at least 50% errors for 2 classes, 33% errors for 3 classes etc., and would thus be a poor performer at the equilibrium point. Of course, it might perform quite well for probability densities other than those giving the equilibrium point. Why design the recognizer for the worst possible p 's (overlapping densities) when these give unacceptable performance? It may be better to restrict the p 's to those given Q_{∞} less than a certain fixed upper bound. It may also be desirable to restrict the \mathcal{R} 's so that the physical realization will be reasonably economical. Note that the class of \mathcal{R} 's will usually include recognizer which "know" the correct value of the statistical parameters before estimation, but there are no grounds to eliminate these, and they cause no difficulty when the game theory approach is used since they do poorly against other p 's. These R 's which know the correct parameter values do cause difficulty if some formal method of optimization such as the calculus of variations is used to find the best estimation for fixed p .

4.1 Pattern detection

The sections above treat the problem of deciding which of k classes a pattern belongs to. The present sections are concerned with the problem of pattern detection. This might be considered as pattern recognition with $k = 2$, the first class being "pattern present", the second class being "no pattern - only normal background present". Several other differences between classification and detection might be noted. Often a detector is left running continuously in time and hence it observes overlapping samples. The criterion of performance for a detector is often of the Neyman-Pearson type, ie. the probability of missing the desired pattern when it occurs is to be minimized while keeping constant the "false alarm probability", the probability of saying a pattern is there when it is not. This is in contrast to the usual symmetrical classification criterion, that of minimizing the average probability of misclassification. If, in detection, there is one or more types of disturbance from the normal background, the problem may be stated as that of classifying into $k-1$ disturbance classes (only one of which is the desired disturbance), and 1 background class. This comes back to another type of pattern recognition in which no decision is made unless the maximum score function exceeds a fixed threshold, eg. 26 letters and "ambiguous letter"⁽⁹⁾. As an example of the detection process, consider a seismic waveform as the pattern with the following classes: 1) atomic explosion, 2) chemical explosion, 3) natural earthquake, 4) natural background - ie. all else. The detector is desired to give a warning when class 1, 2 or 3 has occurred, and then it is expected to decide whether the disturbance is of class 1 or not. Another difference between detection and classification is that in the former the event to be detected might out of control of the experimenter and too rare to gather good statistics on, while in the latter controlled calibration experiments can usually be made. In the worst case it may be desired to detect an event which has not yet occurred, ie. to detect any unusual deviation from the norm.

4.2 Matched filters

The methods used by workers in the field of pattern recognition seem entirely different from those used by communication engineers, yet the problems are very similar examples of statistical detection. In the former field it is common to use computer programs with multivariate gaussian matrix methods, while in the latter field one thinks of matched RLC filters with additive noise. It is the purpose of this section to show the connections between these two approaches, well known to some people; and then to apply some results in multichannel filter design, not published mainly, to pattern detection. The resulting system would use simple analog devices working in real time, and could be operated continuously for pattern detection at unknown times.

The problem of pattern recognition can be described very briefly as follows: A pattern generator G is given the class j and this results in the pattern ψ . The recognizer R looks at the pattern ψ and decides on a class h .



The problem of detection is a special case where j and h take on only two values, signal present or signal not present. The difficulties arise from several sources. First, while j is usually a selection of one of a small set of integers, ψ can be quite a complicated pattern. Second, if the same choice j is repeated several times, the pattern ψ may be different each time. Third, while it might be said formally that there exists a conditional probability of ψ given j , $p(\psi | j)$, the functional form of this distribution are almost never known beforehand. Fourth, even if the transformation $R(\psi \rightarrow h)$ could be calculated on paper in some optimum manner, its physical realization might be far too expensive.

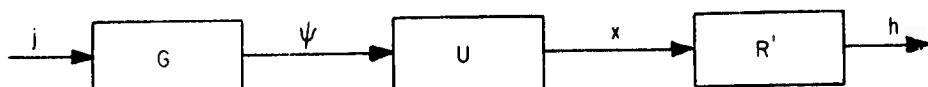
A formal solution to the problem of finding the functional relation $h = Z(\psi)$ of the pattern recognizer R above can be found by using Bayes' rule as shown in the Appendix 1. Thus, $Z(\psi)$ takes on the values h_0 where h_0 is the value of h which maximizes

$$g(h | \psi) = \frac{p_h p(\psi | h)}{g\psi} \quad (54)$$

$$\text{where } g\psi = \sum_j p_j p(\psi | j)$$

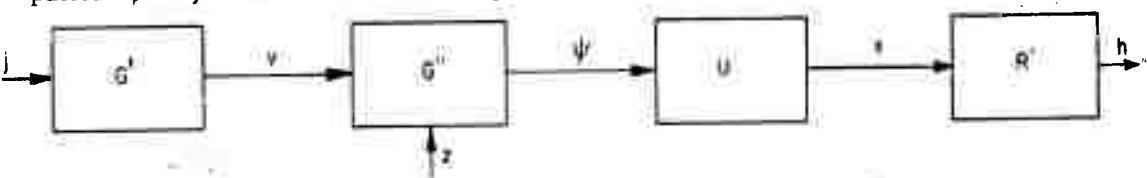
If the above expression has several equal values of h which attain the maximum, an arbitrary rule can be followed in picking one of them. As shown in Appendix 1, this definition of $Z(\psi)$ minimizes the probability of misclassification. The actual calculation is much more complicated than it might appear from this description because ψ may be a vector of many components.

The problem of designing box R in an optimum manner can be simplified by generalizing a procedure of Zadeh and Ragazzini⁽¹⁾. Suppose that a non-singular transformation U is added to the block diagram shown above:



It is obvious that if R' is optimum in deciding the cause j of the pattern x , then UR' is optimum in deciding the cause j from the pattern ψ , ie. UR' is an optimum design for R . The box U can be chosen so as to make the design of R' as easy as possible, as long as it allows ψ to be calculated from x . The original suggestion was to make U uncorrelate the components of ψ occur, or if the components of ψ occur sequentially, to make U flatten the spectrum of one of the pattern classes (noise).

A final modification of the system will be made before considering the actual processes in more detail. The symbol ψ can represent a very complicated vector in many pattern recognition problems, the design of R' is certainly easier if x is a simple vector, but how can U reduce the dimensionability of a signal and still be non-singular? The answer lies in the fact that much of the complexity of ψ is due to random perturbations from an ideal pattern. Suppose that the system is to recognize hand-printed letters; most of the difficulty comes from the fact that each time a particular letter is printed it is slightly different from other samples, even from the same person. In the case of phoneme recognition, each time a "long a" is voiced the waveform does not trace exactly the same values, even with the same speaker using the same word. The pattern ψ may be said to be due to a prototype v and random disturbances z :



The requirements on U may now be eased by asking only that v can be calculated from x , not that v and z can be. Since there are fewer unperturbed patterns (prototypes) than perturbed patterns, v can be of much lower dimensionality than ψ , so can x . It may be convenient to leave some of the patterns randomness in what is called v , and that x may contain more information than just that necessary to calculate v . Note the advantages which ensue from the block "being representable by a group of transformation, as described in Reference 13.

A basic procedure for designing a pattern recognizer or detector will now be as follows: First, it must be found what class of optimum decision box R' can be built. This class may be limited by complexity of apparatus required, by the availability of useful theories of design, and by the information needed on the statistics of the x . Next, the box U must be designed bearing in mind the need to supply a suitable signal to R' ; and the need for preserving the essential information in v .

In order to illustrate the design procedure for the decision box R' , first consider the case where all r components of the vector x are statistically independent (for a fixed j), and where each x_i is normally distributed with the mean depending on j

and with equal variances. If Λ_j is defined as the natural log of $p_j p(x|j)$

$$\Lambda_j(x_1, x_2, \dots, x_r; j) = \ln p_j + \ln p(x_1, x_2, \dots, x_r | j) \quad (55)$$

$$\Lambda_j(x_1, \dots, j) = \ln p_j - r \ln (\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^r [x_i - m_{ji}]^2 \quad (56)$$

where σ is the variance of the x 's

m_{ji} is the mean of x_i from class j

If all of the classes are equally probable then p_j is a constant and the decision function $Z(j)$, which is defined as the value of j which maximizes $p_j p(x|j)$, is given by the value of j which minimizes

$$\sum_{i=1}^n [x_i - m_{ji}]^2 \quad (57)$$

This is recognized as the Euclidean distance from the vector \underline{x} to the vector \underline{m}_j , so that the decision box simply picks the closest class centrum in the feature space $\{\underline{x}\}$. In the case of detection, there are just two classes: noise where $j = 0$ and signal + noise where $j = 1$. Since the noise usually has a zero mean, the question is simply whether the first or second of the quantities

$$g_1 = \sum_{i=1}^r x_i^2 - 2 \sum_{i=1}^r m_{1j} x_i + \sum_{i=1}^r m_{1j}^2 \quad \left. \right\} \quad (58)$$

$$g_0 = \sum_{i=1}^r x_i^2$$

is the smaller. There is a signal present if $g_1 < g_0$, that is if

$$\sum_{i=1}^r m_{1j} x_i > \frac{1}{2} \sum_{i=1}^r m_{1j}^2 \quad (59)$$

Ordinarily, in pattern recognition a calculation must be made for each of the classes involved, or a device must be made to calculate the probability of each class from the x 's. This can be generalized so that $k-1$ devices (or computer subroutines) can make a decision involving k classes, but the loss of symmetry would probably outweigh any

simplification gained in most cases. Returning to the last equation, the recognizer is simply linearly correlating the received set of x^i 's with the signal m and requiring that this correlation be at least half the correlation of m with itself. If the x^i 's occur sequentially in time, then the last equation describes the action of a linear electrical filter whose impulse response is $m_{1, r-i+1}$, the familiar North⁽²⁾ filter of radar usage.

A generalization of the above process can lead to the filter designed by Dwork⁽³⁾ to find a pulse in non-white noise. Suppose now that each class has a correlated normal distribution:

$$p(x_1, x_2, \dots, x_r | j) = \frac{1}{\det D^{\frac{1}{2}}} e^{-\sum_{i, \ell=1}^n D_{i\ell} (x_i - m_{ji})(x_\ell - m_{j\ell})} \frac{(2\pi)^{\frac{r}{2}}}{(2\pi)^{\frac{n}{2}}} \quad (60)$$

Ordinarily, both the means m_i and the $D_{i\ell}$ (which are elements of the reciprocal of the covariance matrix) will be functions of j . However, if signal and noise are added linearly, the $D_{i\ell}$ will be independent of j . If the noise has zero means, the question of signal + noise or noise now depends on the smaller of:

$$g_1 = \sum_{i, \ell=1}^r D_{i\ell} (x_i - m_i)(x_\ell - m_\ell) \quad g_0 = \sum_{i, \ell=1}^r D_{i\ell} x_i x_\ell \quad (61)$$

Again, a signal is present if $g_1 < g_0$, that is if

$$\sum_{i, \ell=1}^r D_{i\ell} m_i x_\ell > \frac{1}{2} \sum_{i, \ell=1}^r D_{i\ell} m_i m_\ell \quad (62)$$

If the x_i occur sequentially, this describes (by a double convolution) the passage of the x_i thru two linear filters in tandem:

$$\sum_{\ell=1}^r D_{i\ell} x_\ell = x_i^i \quad (63)$$

$$\sum_{i=1}^r m_i x_i' = x''$$

The first is a "spectrum-flattening filter" whose impulse response is $D_{i, \ell-r+1}$. Statistically speaking, this first filter decorrelates the variables x_i . The second filter is again a "matched filter" for the signal m_i . Incidentally, the first filter can be looked on as an example of box U as described in the last section, and is in fact the same as the employed by Zadeh and Ragazzini.

The general correlated normal distribution for the x leads to quadratic (rather than linear) decision box. The most probable class $j = Z'(x)$ is that value of j which minimizes

$$g_j = \sum_{i, \ell=1}^r D_{j i \ell} \left[x_i - m_{j i} \right] \left[x_\ell - m_{j \ell} \right] - \frac{1}{2} \ell \ln |\det D_{j i \ell}| \quad (64)$$

This might be regarded as a generalized distance from the unknown to be classified, x , to the centrum of the j class, m_j . If the covariances ϵ are small compared to the variances σ^2 :

$$C_{i \ell} = \begin{cases} \sigma_i^2 & i = \ell \\ \epsilon_{i \ell} & i \neq \ell \end{cases} \quad (65)$$

then an approximate formula for the inverse is

$$D_{i \ell}^{-1} \approx \begin{cases} \sigma_i^{-2} & i = \ell \\ -\sigma_i^{-2} \sigma_\ell^{-2} \epsilon_{i \ell} & i \neq \ell \end{cases} \quad (66)$$

The determinant will be approximated to the same order, by considering only the product of terms on the main diagonal.

$$g_j \approx \sum_{i=1}^r \frac{\left[x_i - m_{j i} \right]^2}{\sigma_{j i}^2} - \sum_{i \neq \ell}^r \epsilon_{j i \ell} \frac{x_i - m_{j i}}{\sigma_{j i}^2} \frac{x_\ell - m_{j \ell}}{\sigma_{j \ell}^2} - \sum_{i=1}^r \ln \sigma_{j i} \quad (67)$$

This form permits first order perturbations of the uncorrelated filter, and it avoids the necessity for calculating the inverse of the covariance matrix.

A matrix generalization of the North-Dwork filter has been given by D. C. Youla⁽³²⁾. Suppose that there are μ channels with time functions $x_1(t), x_2(t) \dots x_\mu(t)$. Let $w_\phi(t)$ be the desired signals (all due to a single disturbance) at time t_0 with spectral $W_\phi(f)$, and let the Fourier transform of the crosscorrelation function between the background noises in channels ϕ and ν be $\Omega_{\phi\nu}(f)$. The filter is to have μ inputs one output, each having a voltage function of time associated with it. The transfer functions $R_\nu(f)$ between the inputs $\nu = 1, 2 \dots \mu$ and the output should then satisfy the following simultaneous algebraic equations in order to maximize the ratio of peak signal to RMS noise at the filter output:

$$W_\phi^*(f) = \sum_{\nu=1}^{\mu} \Omega_{\phi\nu}(f) R_\nu(f) \quad \phi = 1, 2 \dots \mu \quad (67)$$

If $\mu = 1$ this reduces to the Dwork filter⁽³⁾:

$$R_1(f) = \frac{W_1^*(f)}{\Omega_{11}(f)}$$

where Ω_{11} is the power spectrum of the noise. These equations do not generally yield realizable network functions, but allowing a delay in the output permits sufficiently good approximations to be made with materially affecting the results. To obtain the best filter if only a finite delay is allowed, simultaneous Wiener-Hopf integral equations must be solved⁽³²⁾. A diagram of a detection system using multiple matched filters is shown in Fig. 3. The voltage in channel 1 has a + pulse at the time of the disturbance (top of figure) t_0 , channel 2 has a delayed - pulse, channel 3 has a + pulse followed by a - pulse. The filter R combines these in such a way that contributions from all channels add up at time $t_0 + t_d$ and exceed the threshold θ .

The major problem in the design of such a system is to decide on the boxes U_1, U_2, \dots, U_μ . These must provide signals and noises for which a linear filter is reasonably efficient at filtering the signal from noise. Whereas a straight-forward design is the familiar problem of property selection in pattern recognition. Several general methods for selecting properties have been given in a previous report⁽¹³⁾, in Minsky's review article⁽³³⁾, and in Section 6 below. The best method to use if the pattern is of the type used in radio station recognition^(13, 17) is probably a combination of band pass filters and the operator sequences of Selfridge⁽³⁴⁾ and Minsky⁽³³⁾. The operators might be quite similar to those used in the radio station recognition experiments^(19, 23) except that they would have running counts, averages, etc. The carrier signal or even the AGC voltage could not itself be fed to the filter R because they do not repeat accurately

A matrix generalization of the North-Dwork filter has been given by D. C. Youla⁽³²⁾. Suppose that there are μ channels with time functions $x_1(t), x_2(t) \dots x_\mu(t)$. Let $w_\phi(t)$ be the desired signals (all due to a single disturbance) at time t_o with spectral $W_\phi(f)$, and let the Fourier transform of the crosscorrelation function between the background noises in channels ϕ and ν be $\Omega_{\phi\nu}(f)$. The filter is to have μ inputs one output, each having a voltage function of time associated with it. The transfer functions $R_\nu(f)$ between the inputs $\nu = 1, 2 \dots \mu$ and the output should then satisfy the following simultaneous algebraic equations in order to maximize the ratio of peak signal to RMS noise at the filter output:

$$W_\phi^*(f) = \sum_{\nu=1}^{\mu} \Omega_{\phi\nu}(f) R_\nu(f) \quad \phi = 1, 2 \dots \mu \quad (67)$$

If $\mu = 1$ this reduces to the Dwork filter⁽³⁾:

$$R_1(f) = \frac{W_1^*(f)}{\Omega_{11}(f)}$$

where Ω_{11} is the power spectrum of the noise. These equations do not generally yield realizable network functions, but allowing a delay in the output permits sufficiently good approximations to be made with materially affecting the results. To obtain the best filter if only a finite delay is allowed, simultaneous Wiener-Hopf integral equations must be solved⁽³²⁾. A diagram of a detection system using multiple matched filters is shown in Fig. 3. The voltage in channel 1 has a + pulse at the time of the disturbance (top of figure) t_o , channel 2 has a delayed - pulse, channel 3 has a + pulse followed by a - pulse. The filter R combines these in such a way that contributions from all channels add up at time $t_o + t_d$ and exceed the threshold θ .

The major problem in the design of such a system is to decide on the boxes U_1, U_2, \dots, U_μ . These must provide signals and noises for which a linear filter is reasonably efficient at filtering the signal from noise. Whereas a straight-forward design procedure has just been given for R' , no such procedures are known for the U 's. This is the familiar problem of property selection in pattern recognition. Several general methods for selecting properties have been given in a previous report⁽¹³⁾, in Minsky's review article⁽³³⁾, and in Section 6 below. The best method to use if the pattern is of the type used in radio station recognition^(13, 17) is probably a combination of band pass filters and the operator sequences of Selfridge⁽³⁴⁾ and Minsky⁽³³⁾. The operators might be quite similar to those used in the radio station recognition experiments^(19, 23) except that they would have running counts, averages, etc. The carrier signal or even the AGC voltage could not itself be fed to the filter R because they do not repeat accurately

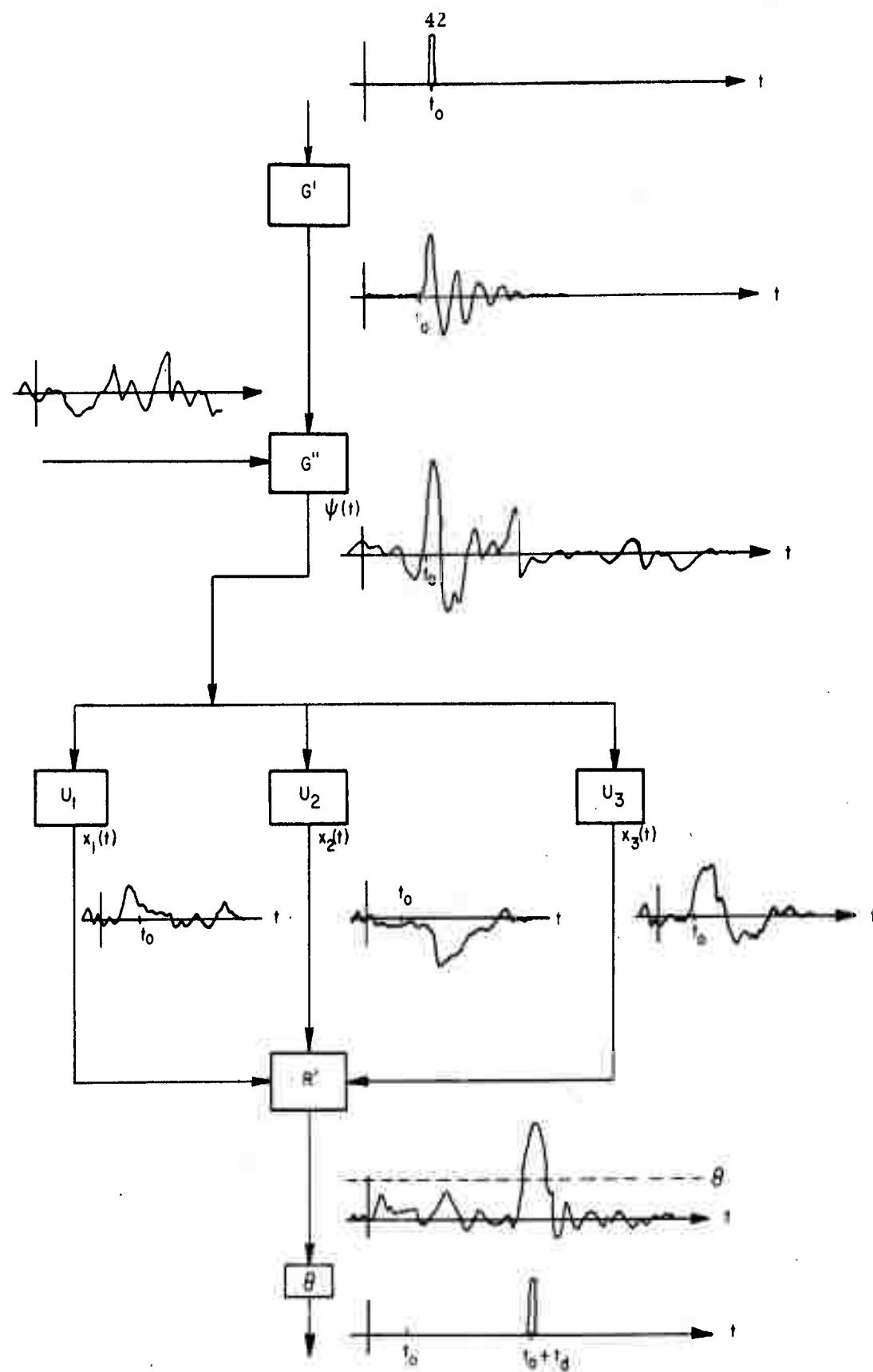


Fig. 3 Pattern Detection With a Multiple-Channel Matched Filter

MRI-19075

enough for successive disturbances. The signals fed to R' would be very low frequency in order to make the desired signal repeatable. The U 's would ideally be selected to make the noises in the various channels have negative correlations where the signals have positive correlations, and vice versa.

5.1 Radio station recognition

The problem of recognizing radio stations from their carrier curves has been chosen as a convenient example to illustrate and test concepts and methods of pattern recognition. Theories of the type given in the sections above should always be supplemented by information obtained from study of the physical mechanism of the pattern generator if the best pattern recognizer is desired. The principal thing desired is to find a suitable set of properties to feed to the decision making part of the recognizer. It is obvious that a property which varies a large amount between classes and which varies a small amount between different samples of the same class is a very good property. This is just saying that it is desirable that the ratio of interclass variance to intraclass variance be large. But note that if the pattern generator is not fixed in time, certain properties may be very useful as controls (see Section 2.7 above) even if they do not vary at all from class to class. Note also that the numerical results of Section 3.1 above show that the statistical independence of the different properties is of great importance, especially if the ratio of interclass variance to intraclass variance is not large. A principle which is useful in designing a pattern recognizer is that the performance cannot be degraded by adding "noisy" or useless properties, although, of course, these properties may not help the performance either⁽¹³⁾. This principle suggests that the designer should be more concerned with not overlooking a mediocre property than with weeding out poor properties.

The problem of interest here is to determine the location of a sine wave radio transmitter in the frequency range 6-20 mc. by study of the received signal. If propagation is by various combinations of reflections (or refractions) from different layers of the ionosphere and reflections from the earth, and if the condition of the ionosphere changes with time, the received signal will not be sinusoidal. Tests which have actually been made have used the carriers of a discrete set of international broadcast stations as the sine wave sources, and since the time variations caused by the ionosphere are usually much slower than any modulation component, the carrier fading curve can easily be isolated, say by the receiver's AGC time constant circuits.

Propagation between two points can usually take place over an infinity of possible paths, but under normal conditions these paths can be grouped, one group corresponding to a single hop involving the F 2 layer, another involving two F 2 hops, another involving a combination of one E layer hop and one F 2 layer hop, etc. Within the group there are many paths quite similar in length representing the scattering or non-specular reflection due to ionospheric inhomogeneities. Balser and Smith⁽²⁴⁾ have recently studied some of the statistical properties of a single mode (group of paths). Their work, together with some older results^(25, 26), suggest the following model for the received signal:

$$G(t) = g_1(t) + g_2(t) + \dots + g_\mu(t)$$

Each of the g 's can be statistically described by giving its first order probability density and its auto correlation function (spectral power density). The different g 's might reasonably be expected to be statistically independent in the sense that if a particular station is received for about a half hour the minute by minute fluctuations of parameters describing the probability densities and autocorrelations about their means are independent. The path lengths are very long in wavelengths and the different paths are reflected from widely separated regions of the ionosphere. The $g(t)$ functions vary more slowly⁽²⁴⁾ than the $G(t)$ function, say up to 0.3 cps as compared to up to 10 cps or more. This is because there can be fast beats between two g 's. A set of properties which might be quite useful in radio station recognition might be the following:

$$\begin{array}{llll} x_1 = f_1 & x_4 = f_2 & \dots & x_{3\mu-2} = f_\mu \\ x_2 = v_1 & x_5 = v_2 & \dots & x_{3\mu-1} = v_\mu \\ x_3 = \delta_1 & x_6 = \delta_2 & \dots & x_{3\mu} = \delta_\mu \end{array}$$

where f_ν is the center frequency of $g_\nu(t)$, v_ν is the variance of $g_\nu(t)$, and δ_ν is the width of the first lobe of the autocorrelation function of $g_\nu(t)$. Several more properties can be obtained from each $g(t)$, for example the parameter "a" of Baker and Smith⁽²⁴⁾, the parameter "m" of Nakagami⁽³⁵⁾, additional parameters describing the shape of the auto correlation function etc. Note that the relative frequency range of $G(t)$ and therefore the f 's (but not their differences) will depend on the local oscillator frequency if a linear detector is used. If an envelope detector is used the frequencies will not depend on the local oscillator but μ will usually be increased and the different g 's will not be independent⁽¹⁷⁾. Note also that there is a problem in how to label the different components in the sequence $g_1(t), g_2(t) \dots$ because if they are numbered by increasing frequency and if the first component fades out all remaining properties will be permuted. If the numbering is in order of decreasing amplitude (v) two nearly-equal components can switch their labels, but this may not cause as much difficulty.

5.2 Radio station recognition results

The results of some more tests of the radio station recognition process are tabulated in Table 7. These tests were designed to study the following effects:

1. The advantage of using a correlated Gaussian distribution compared to the uncorrelated form. An increase in the percent of correct classification can be usually noted if comparable tests are compared (10 and 11 differ only in the inclusion of covariance terms). However, the use of a correlated Gaussian formula with a small sample

means the parameter estimation will be very poor since there are so many covariance terms. If a separate calibration sample is used the percent of correct classification is subject to large sampling errors; if classification is done on the same sample which was used for calibration the result ends to be much too optimistic (e. g. Test 6).

2. The effect of the number of properties used. It has been shown in a previous report⁽¹³⁾ that including too many properties in a pattern recognizer cannot make the expected percent of correct classification lower, although the extra properties may not improve the classification either. From Test 10 to Test 9 there is a slight improvement. As might be expected, the correlated tests show larger improvements, eg. Tests 11 and 14 are comparable and show an improvement from 45% to 78% when the properties are increased from 6 to 12.

3. Controls. The use of a control signal to partially compensate the effects of time-varying statistics has been previously discussed (sections 2.6 and 2.7). Test 11 was made with 6 properties and no control and gave 45% while Test 15 with the same 6 properties and with control gave 69%. There is some indication, Tests 14 and 15, that controls are not as good as an equal number of additional properties.

4. The effect of recognizing on the calibration data is to increase the percent of correct classification. Even if the data happened to be entirely random and independent of the station, if only two patterns from each station are observed they would almost certainly be correctly classified correctly by a recognizer designed from the same two patterns. If many patterns are used the results are not quite so deceptive since a large percent correct does indicate a strong clustering of the sample points. Table 7 indicates a lowered percent correct when the recognition is on separate patterns, but the results are still above those to be expected from a purely random classification. For example, in Test 16 with 8 stations being recognized, if the decision were purely random the percent correct would be expected to be 12.5%, not 34%. The large percent correct in Test 7 may be explained by the fact that only 4 stations were involved.

5. Some other effects have been studied which do not appear in Table 7. A few tests were made with 30 sec. patterns instead of 60 sec. patterns, and the percent correct with 30 sec. patterns was not noticeably lower, but there is too little evidence to come to any definite conclusion. There are theoretical reasons to believe that longer samples will give better results, and this will be tested. Another interesting question is whether the errors which were made would be repeated in the same way if the same recognizer is used on successive 60 sec. patterns on the same day. One test indicates that the error made on a certain day is repeated consistently more often than not, and therefore designing the pattern recognizer to vary from day to day is important. The effect of different properties and different controls (including sunspot numbers

| Test Number | Properties (+ controls) | Stations | Days recognized (calibrated on) | % correct | Remarks |
|---------------------------------|-------------------------|----------|---------------------------------|-----------|------------------|
| <u>UNCORRELATED PROPERTIES</u> | | | | | |
| 10 | 6, | 8 | | | |
| 10 | 6, | 8 | 20 | 32 | |
| 9 | 23 | 8 | 20 | 48 | |
| 27+28 | 23 | 6 | 10 | 51 | 30 sec patterns! |
| Calibration sample different | | | | | sampling error? |
| 7 | 6 | 4 | 3(5) | 67 | |
| <u>CORRELATED PROPERTIES</u> | | | | | |
| 14 | 12 | 8 | 20 | 78 | |
| 11 | 6 | 8 | 20 | 45 | |
| 6 | 6 | 4 | 5 | 95 | sample too small |
| Calibration sample same | | | | | |
| 16 | 12 | 8 | 10(20) | 34 | |
| 8 | 6 | 4 | 3(5) | 33 | |
| <u>CORRELATED WITH CONTROLS</u> | | | | | |
| Calibration sample same | | | | | |
| 15 | 6 + 6 | 8 | 20 | 69 | |
| Calibration sample different | | | | | |
| 17 | 6 + 6 | 7 | 10 (20) | 29 | |

WWV propagation reports, and indices of magnetic activity) have been investigated and will be reported on⁽²³⁾.

5.3 Apparatus

Three pieces of apparatus have been designed and built on this contract and in connection with the previous contract: an analog-to-digital converter⁽¹⁸⁾, a subsonic spectrum analyzer⁽²¹⁾, and a stable RF source⁽³⁶⁾. A block diagram of the spectrum analyzer is shown in Fig. 4. A photograph of the apparatus is shown in Fig. 5, and some waveforms are shown in Figs. 6, 7. The apparatus is essentially finished and all blocks have been tested, but an overall test of the closed loop has not yet been made.

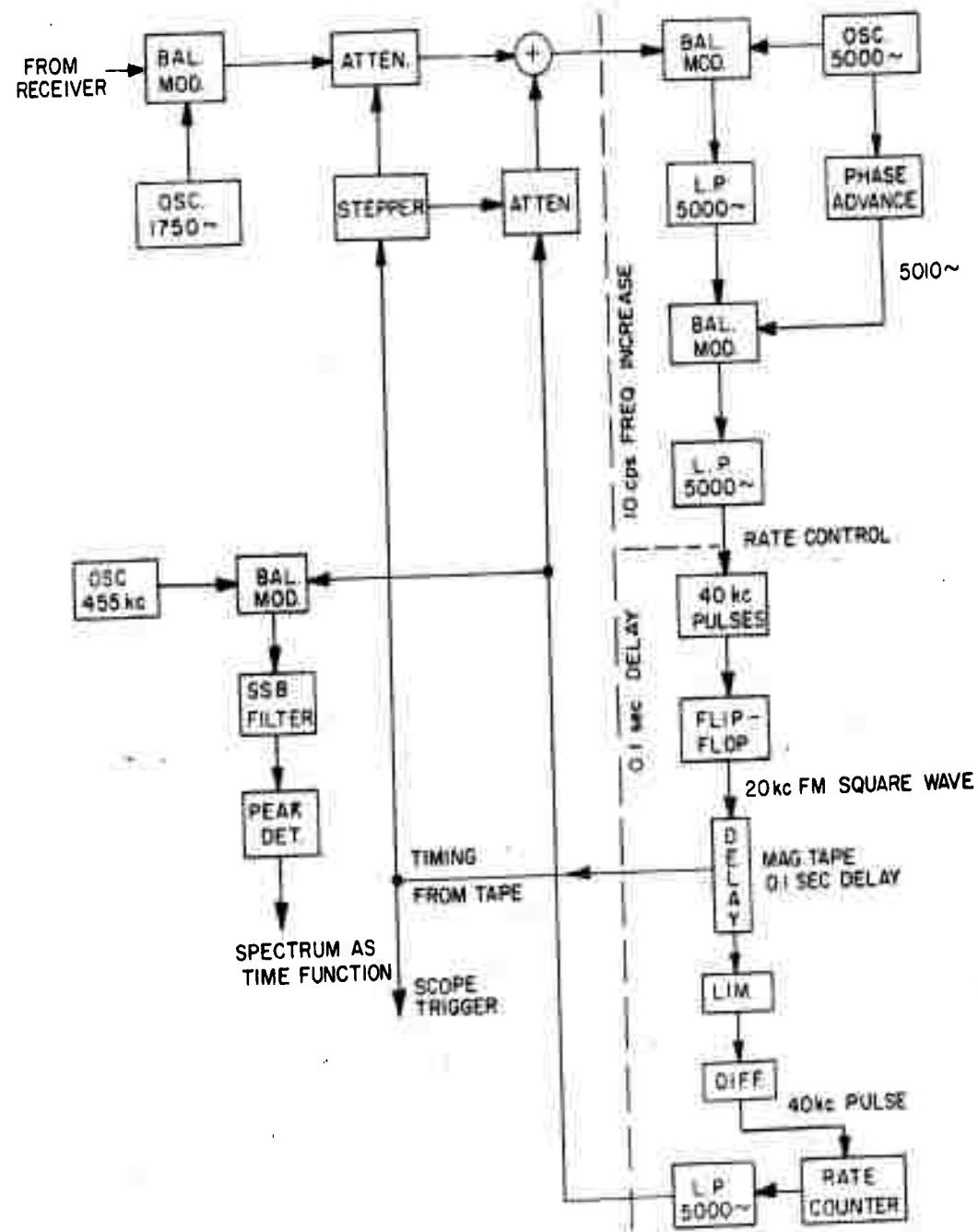
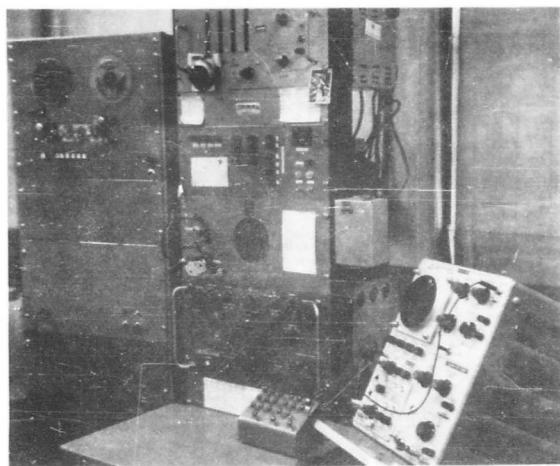


Fig 4 Subsonic Spectrum Analyzer

Spectrum Analyzer shown
with A-D Converter and Radio
Receiver



Magnetic Tape Delay Unit



10 cps Frequency Shifter and
FM Modulator



Stepped Attenuator



FM Demodulator and Input
Modulator

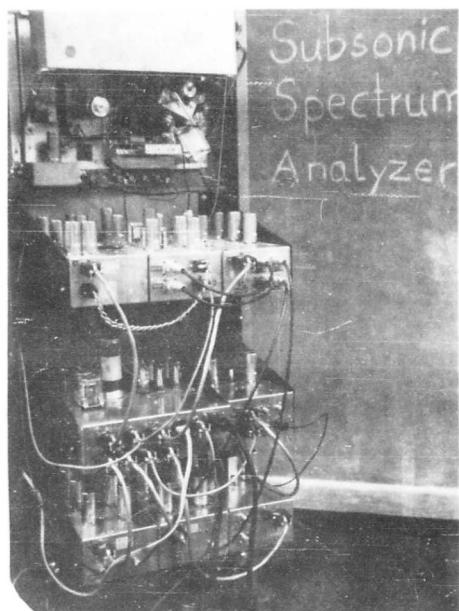
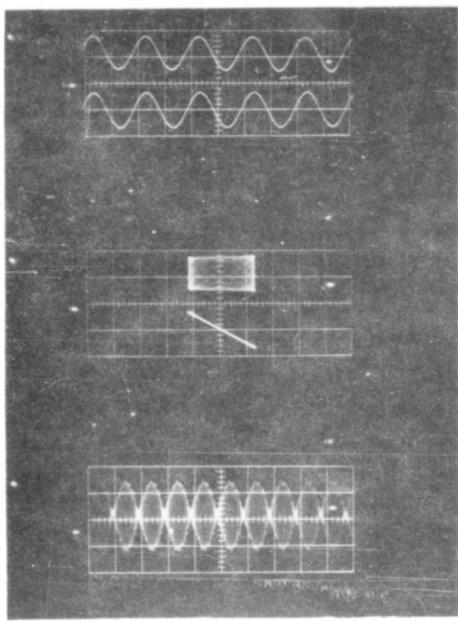


Fig. 5 Spectrum Analyzer

I

1. Input Signal 2500 cycles

- .5V / Major division
- .2 msec. / Maj. Div.

2. Frequency Translator Output

- .5V / Major Division
- .2 msec. / Maj. Div.

3. Lissajous Fig. Input versa Input

- .5V / Maj. Div.

4. Lissajous Fig. Input versa Output

- .5V / Maj. Div.

5. Algebraic Sum of Input + Output

- .5V / Maj. Div.
- .1 sec / Maj. Div.

II

1. Input Signal 1800 cycles

- .5V / Major Division
- .2 msec / Maj. Div.

2. Frequency Translator Output

- .5V / Major Division
- .2 msec / Maj. Div.

3. Lissajous Fig. Input versa Input

- .5V / Maj. Div.

4. Lissajous Fig. Input versa Output

- .5V / Maj. Div.

5. Algebraic Sum of Input + Output

- .5V / Maj. Div.
- .1 sec / Maj. Div.

Fig. 6 Waveforms from Frequency Shifter

1. Frequency Translator Output, (Input 2500 cycles)

- .5V/Maj. Div.
- .1 msec/Maj. Div.

2. FM-Modulator Output

- 20V/Maj. Div.

3. FM-Modulator Output

- 20V/Maj. Div.

4. Readhead Output

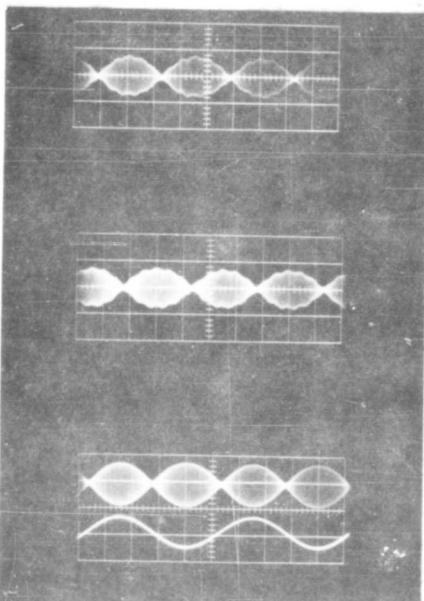
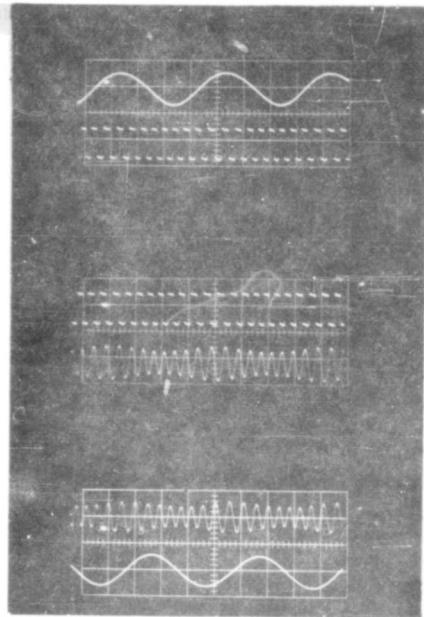
- .1V/Maj. Div.

5. Readhead Output

- .1V/Maj. Div.

6. FM-Denominator-Output

- .5V/Maj. Div.
- .1 msec/Maj. Div.

1. 10 Cycles Sinewave

- IV/Maj. Div. .2 Sec/Maj. Div.

2. Input Modulator Output

- .5V/Maj. Div.
- .2 Sec. /Maj. Div.

3. Frequency Translator Output

- .5V/Maj. Div.
- .2 sec/Maj. Div.

4. Delay-line Output

- .5V/Maj. Div.
- .2 sec/Maj. Div.

Fig. 7 Waveforms from Delay Unit and Input Modulator

6.1 General design procedure for pattern recognizers

No attempt has been made in the above sections or in the previous report⁽¹³⁾ to cover all methods of pattern recognition which have been suggested in the literature. Most of the published attempts at pattern recognition have been specifically tailored to one application and contribute little to a generalized theory of pattern recognition. This is not to say that a very good pattern recognizer cannot be made by someone who knows nothing of statistics and pattern recognition theory, but who knows much about the particular pattern generating mechanism or who can intuitively see how humans are performing the same recognition. Indeed, it would be foolish not to use all knowledge which is available concerning the particular field in which the pattern recognition is to take place. All special tricks which can be thought of should be employed. The complete theory should desirable be general enough to include all special cases. However, there are interesting areas where the more formal theory is the only guide: Humans have no experience in recognizing the pattern, a complete physical theory of the pattern generator is impractical, and the pattern itself is too complex for simple enumerative techniques. It is not felt that a complete formal theory of pattern recognition is yet at hand. The following is meant to be a checklist of idea which might serve as components in such a theory. More details on each topic can be found in the references cited. The reviews of Minsky⁽³³⁾, Hawkins⁽⁴⁴⁾, Sebestyn⁽⁶⁹⁾, Weisz et al⁽⁴³⁾, and the Bionics Symposium Report⁽⁵²⁾ should provide references to most of the work in pattern recognition up to 1961.

A. Property selection methods

1. Decision tree: Morse code recognizer⁽⁴⁷⁾, property lists and characters⁽³³⁾, more properties for difficult pairs of classes sequential decisions.
2. Initial transformation: autocorrelation⁽⁴⁶⁾, pre-recognition cleanup⁽⁵⁰⁾, decorrelation⁽⁴⁹⁾, segmentation of handwritten letters⁽⁶⁸⁾, normalization of size⁽³³⁾, sampling of a time function⁽³⁰⁾, detection of RF signals⁽¹⁷⁾, clustering⁽⁶⁹⁾.
3. Search and learning: pandemonium⁽⁴⁵⁾, operator sequences⁽³⁴⁾, n-grams in language recognition⁽¹³⁾, hill climbing⁽³³⁾.
4. Random selection^(37, 38), random networks⁽⁴⁴⁾, perceptron⁽⁵⁴⁾.
5. Coding theory⁽⁵⁶⁾: parity checks⁽⁴⁴⁾, group codes⁽¹³⁾.
6. Invariants to distorting transformations: to groups of transformations^(39, 40, 41), topological invariants⁽⁴²⁾, transformation not a group⁽¹³⁾.
7. Physical model of pattern generator as a guide: radio station recognition⁽¹⁷⁾, speech recognition⁽⁵³⁾, template matching⁽³³⁾.

8. Imitate biological processes^(43, 52), spatial computers⁽⁴⁸⁾, retinal structure⁽⁵¹⁾, search for distinctive features "by eye"⁽⁶⁷⁾.

B. Decision making in property space.

1. Non-statistical methods: logical⁽⁵⁵⁾, linear boundaries⁽⁹⁾, convex cones⁽²⁷⁾, potential fields⁽¹³⁾, distance comparisons⁽¹³⁾, correlation⁽¹³⁾, quadratic boundaries⁽⁵⁷⁾, polynomial boundaries⁽⁶⁹⁾.

2. Choice of probability density function: uncorrelated Gaussian, correlated Gaussian with equal covariance matrices^(8, 10), ditto with separate covariance matrix for each class^(13, 23), discriminating functions^(13, 58, 66), spherically symmetrical densities⁽⁶⁰⁾, Poisson distributions.

3. Parameter estimation: classical methods⁽⁵⁾, hypothesis testing⁽⁷⁾, no explicit estimation (see Section 2 of this report).

4. Non-stationary conditions (see Section 2 of this report), prediction^(61, 62).

5. Miscellaneous topics: decision theory⁽⁵⁹⁾, game theory^(15, 16), context correction after recognition after recognition⁽⁶⁵⁾, "no decision" regions⁽⁹⁾, use of human link, adaptive filters^(63, 64).

6. Testing: partition of a fixed sample between calibration and recognition⁽⁹⁾, misleading to recognize calibration points, is population really Gaussian⁽⁸⁾.

Appendix - Bayes' Rule

Bayes' rule can be applied to almost any situation where an unseen cause is to be inferred from an observed effect. The joint probability P of a cause c ($c = 1, 2 \dots n$) and effect e ($e \in E$, some probability space) can be written:

$$P(c, e) = p_c p(e|c)$$

where p_c = a-priori probability of cause c

$p(e|c)$ = conditional probability of effect e given cause c

The conditional probability of the cause c given the effect e , $q(c|e)$, is obtained by dividing P by the probability of the effect e :

$$q(c|e) = \frac{P(c, e)}{q_e} = \frac{p_c p(e|c)}{\sum_k p_k p(e|k)} \quad (A 1)$$

Bayes' rule is to select the most probable cause given the effect e , that is to maximize the above expression as a function of c . This defines a function of e , say $\hat{c}(e)$, taking on values $1, 2 \dots n$. The space E is thereby divided into n regions $R_1, R_2, \dots R_n$, where $e \in R_k$ if and only if $\hat{c}(e) = k$.

Note that since the denominator does not depend on c , only the numerator need be considered in the maximization. (If p_c is either unknown or a constant, then $p(e|c)$ will be maximized as a function of c resulting in a "maximum likelihood" decision.)

The probability of making an incorrect decision by following Bayes' rule is given by:

$$Q = 1 - \sum_{k=1}^n p_k \int_{R_k} p(e|k) de \quad (A 2)$$

$$Q = 1 - \int_E p_{\hat{c}(e)} p(e|\hat{c}(e)) de \quad (A 3)$$

Consideration of the definition of R_k and $\hat{c}(e)$ shows these forms to be equivalent, and it is obvious from the second that Q will be minimized by following Bayes' rule.

Some very useful variants of Bayes' rule can easily be developed. Suppose there are two causes c_1 and c_2 , but that only c_1 is to be inferred. The joint probability of cause and effect is now $P(c_1, c_2, e)$ while the probability of the desired cause given the effect is

$$q(c_1 | e) = \frac{\sum_{c_2} P(c_1, c_2, e)}{q_e} = \frac{\sum_{c_2} p_{c_1, c_2} p(e | c_1, c_2)}{q_e}$$

The joint probability of the causes is $p_{c_1} p_{c_2 | c_1}$, therefore

$$q(c_1 | e) = \frac{p_{c_1} \sum_{c_2} p_{c_2 | c_1} p(e | c_1, c_2)}{q_e}$$

However, by using the fact that any probability formula remains value if the same condition is put in all probabilities:

$$p(e | c_1) = \sum_{c_2} p_{c_2 | c_1} p(e | c_1, c_2)$$

the following is obtained

$$q(c_1 | e) = \frac{p_{c_1} p(e | c_1)}{q_e}$$

This is the same as completely ignoring the undesired cause in all formulas from the beginning.

The above may be obvious, but it gives some formulas useful in the following situation: There are two causes c_1 and c_2 and two effects e_1 and e_2 . Effect e_1 is influenced by both causes, but effect e_2 is influenced only by cause c_2 and is independent of cause c_1 . Only cause c_1 is desired. If the last paragraph is reread it will be found to be perfectly valid if e is replaced by e_1, e_2 , since no mention was made as to whether the different parts of e might or might not be independent of c_1 or c_2 . The last equation becomes

$$q(c_1 | e_1, e_2) = \frac{p_{c_1} p(e_1, e_2 | c_1)}{q_{e_1, e_2}} \quad (A 4)$$

The data e_2 might be called a "control" or "side information" since, although it does not depend on the desired cause c_1 , it does depend on c_2 which has an undesired influence on e_1 . Such side information should be included with the other effects. If this

situation is looked on from another angle, it may be supposed the e_2 should be used to infer something about c_2 , which could in turn modify e_1 and so improve the estimate of c_1 . This approach would not only be more complicated, but it could not be more effective than the one given. Some other forms for the above equation can be obtained as follows:

$$p(e_1, e_2 | c_1) = p'(e_2 | c_1) \mu(e_1 | c_1, e_2)$$

but since e_2 does not depend on c_1 :

$$p'(e_2 | c_1) = q_{e_2}$$

Combining these equations:

$$q(c_1 | e_1, e_2) = \frac{p_{c_1} \mu(e_1 | c_1, e_2)}{q_{e_1} | e_2} \quad (A 5)$$

The meaning of the variables appearing in the above, and the use to which the formulas might be put, will perhaps be more clearly fixed in mind if a couple of examples are given. Example 1: Consider speech recognition and let c_1 be the desire to utter a certain phoneme, let c_2 be some physical characteristic of the speaker, say his weight. Now a person's weight will not be influenced by which phoneme he tries to utter, but it is well known that present speech recognizers work well on the person who calibrated it and poorly on other persons. Perhaps the weight will at least give some clue as to the general pitch of the speaker, and the recognizer may be able to do better knowing this general pitch. Example 2: Suppose the usual Bayes' rule situation, a-priori probabilities unknown, conditional probabilities known, is reversed. In a given cause-effect situation the a-priori probabilities are known but the "black box" probability (of the outcome given the input) is not known. Identify the variables as follows: c_1 the input to be guessed, c_2 the particular "black box" in use, e_1 the observed outputs due to the input which is to be guessed, and e_2 the observed outputs for a calibration run during which known inputs are fed to the "black box". Since the calibration outcomes cannot depend on the selection of the unknown input (which occurs later in time), and since the particular "black box" in use is of no interest, the usual calibrated pattern recognizer fits this Bayes' rule extension.

REFERENCES

1. L. A. Zadeh and J. R. Ragazzini, "Optimum Filters for the Detection of Signals in Noise", Proc. of IRE, Vol. 40, pp. 1223-31, Oct. 1952.
2. J. L. Lawson and G. E. Uhlenbeck, "Threshold Signals", McGraw Hill, New York 1950. See page 209.
3. B. M. Dwork, "Detection of a Pulse Superimposed on Fluctuation Noise", Proc. of IRE, Vol. 38, pp. 771-4, July 1950.
4. W. B. Davenport and W. L. Root, "An Introduction to the Theory of Random Signals and Noise", McGraw Hill, New York, 1958.
5. H. Cramér, "Mathematical Methods of Statistics", Princeton U. Press, Princeton, 1946.
6. P. W. Cooper, "Classification by Statistical Methods (Pattern recognition)", Melpar Technical Note No. 61/2, April 1961.
7. C. R. Rao, "Statistical Inference Applied to Classificatory Problems", Chapter 8 in "Advanced Statistical Methods in Biometric Research", John Wiley and Sons, New York 1952.
8. T. W. Anderson, "Classification of Observations", Chapter 6 in "An Introduction to Multivariate Statistical Analysis", John Wiley and Sons, New York, 1958.
9. W. H. Highleyman, "Linear Decision Functions, With Application to Pattern Recognition", DEE Dissertation, Polytechnic Inst. of Brooklyn, June 1961. - also Proc. of IRE, Vol. 50, June 1962.
10. T. Marill and D. M. Green, "Statistical Recognition Functions and the Design of Pattern Recognizers", IRE Trans, Vol. EC9, pp. 472-7, Dec. 1960.
11. E. Jahnke and F. Emde, "Tables of Functions with Formulas and Curves", Dover, New York, 1945.
12. K. Pearson, "Tables for Statisticians and Biometricalians, Part 2", Cambridge U. Press, 1931.
13. A. E. Laemmel, "Machine Learning Processes Applied to Pattern Recognition", Final Report No. PIBMRI 981-61 (AFCRL 62-6), Polytechnic Inst. of Brooklyn, 14 Dec. 1961.
14. S. O. Rice, "Communication in the Presence of Noise-Probability of Error for Two Encoding Schemes", Bell Sys. Tech.J., Vol. 29, pp. 60-93, 1950.
15. J. von Neumann and O. Morgenstern, "Theory of Games and Economic Behavior", Princeton U. Press, 1944.

16. R. D. Luce and H. Raiffa, "Games and Decisions, Introduction and Critical Survey", John Wiley and Sons, New York, 1958.
17. A. E. Laemmel, "Location of HF Radio Stations From Their Fading Characteristics", Report RIBMRI-962-61, Scientific Report No. 1 (AFCRL 62-79), Polytechnic Inst. of Brooklyn, 19 Oct. 1961.
18. W. A. Bellmer, "Description and Manual of Operation for Analog to Digital Converter", Report PIBMRI-978-61 (AFCRL-62-99), ibid., 29 Jan. 1962.
19. B. Rudner and W. A. Bellmer, "Radio Station Recognition Computer Programs and Test Results", Report PIBMRI-979-61, Sci. Rpt. No. 2, (AFCRL-62-564) ibid., 12 Jan. 1962.
20. B. Rudner, "Algebraic Computer Learning Experiment Report PIBMRI-980-61, Sci. Rpt., No. 3 (AFCRL-62-565), ibid., 17 Jan. 1962.
21. T. Bially, "Theory and Design of a Subsonic Spectrum Analyzer", Report PIBMRI-1013-62, Sci. Rpt. No. 1 (AFCRL-62-368) ibid., 26 April 1962.
22. A. E. Laemmel, "Measurement of Ionospheric Paths as Time-Varying Network", Report PIBMRI-1102-62, Sci. Rpt. No. 2 ibid., 13 Dec. 1962.
23. B. Rudner and W. B. Bellmer, "Experiments in Radio Station Recognition", PIBMRI- 1135-63, Sci. Rpt. No. 3.
24. M. Balser and W. B. Smith, "Some Statistical Properties of Pulsed Oblique HF Ionospheric Transmissions", Lincoln Laboratory Preprint, 1962.
25. V. Agy, K. Davis and R. Salaman, "An Atlas of Oblique-Incidence Ionograms", NBS Tech. Note No. 31, Nov. 1959.
26. L. B. Arguimbau et al, "Transatlantic Frequency-Modulation Experiments", Tech. Rpt. 278, MIT Research Lab. of Electronics, 20 Sept. 1954.
27. R. M. Blood, "Inductive Processes for Factor Selection in a Recognition Model", Vitro Laboratories, West Orange, New Jersey.
28. H. Sherman, "A Quasi-Topological Method for Machine Recognition of Line Patterns", Proc. of Internatl. Conf. on Information Processing, UNESCO, Paris France 1959.
29. C. E. Shannon, "A Mathematical Theory of Communication", Bell System Tech. Journal, Vol. 27, pp. 379-423, 623-56, 1948.
30. C. E. Shannon, "Communication in the Presence of Noise", Proc. IRE, Vol. 37, pp. 10-21, Jan. 1949.

31. R. M. Fano, "Transmission of Information - A Statistical Theory of Communication", John Wiley and Sons, New York, 1961.
32. D. C. Youla, "The Theory and Design of Multiple Channel Matched Filters", report for the Atlantic Research Corp., 25 June 1959.
33. M. Minsky, "Steps Toward Artificial Intelligence", Proc. of IRE, Vol. 49, pp. 8-30, Jan. 1961.
34. O. G. Selfridge, "Pattern Recognition and Modern Computers", Proc. of Western Joint Computer Conf., March 1955.
35. M. Nakagarni, "The m-Distribution - A General Formula of Intensity Distribution of Rapid Fading", in "Statistical Methods in Radio Wave Propagation", ed. by W. C. Hoffman, Permagon, New York 1960.
36. P. Boros, "Stable Radio Frequency Source", MEE Thesis, Polytechnic Inst. of Brooklyn, June 1962.
37. W. W. Bledsoe and I. Browning, "Pattern Recognition and Reading by Machine", Proc. of EJCC, pp. 225-32, Dec. 1959.
38. G. P. Steck, "Stochastic Model for the Browning-Bledsoe Pattern Recognition Scheme", IRE Trans., Vol. EC11, pp. 274-82, April 1962.
39. W. Pitts and W. S. McCulloch, "How We Know Universals", Bull. Math. Biophysics, Vol. 9, pp. 127-47, Sept. 1947.
40. L. P. Eisenhart, "Continuous Groups of Transformations", Dover, New York, 1961.
41. H. Weyl, "The Classical Groups, Their Invariants and Representations", Princeton, 1946.
42. H. Sherman, "A Quasi-Topological Method for Machine Recognition of line Patterns", Proc. of Conf. on Info. Processing, UNESCO Paris 1959.
43. A. Z. Weisz, J. C. R. Licklider, J. A. Swets, and J. P. Wilson, "Human Pattern Recognition Procedures is Related to Military Recognition Problems", Bolt Beranck and Newman Inc. Report No. 939, 15 June 1962.
44. J. K. Hawkins, "Self Organizing Systems - A Review and a Commentary", Proc. IRE, Vol. 49, pp. 31-48, Jan. 1961.
45. O. G. Selfridge, "Pandemonium: A Paradigm for Learning", in "Mechanisation of Thought Processes", Her Majesty's Stationery Office, London 1959.
46. L. P. Horwitz and G. L. Shelton Jr., "Pattern Recognition Using Autocorrelation", Proc. IRE, Vol. 49, pp. 175-185, Jan. 1961.

47. B. Gold, "Machine Recognition of Hand-Sent Morse Code", IRE Trans. Vol. 1T5, pp. 17-24, March 1959.
48. S. H. Unger, "A Computer Oriented Toward Spatial Problems", Proc. IRE, Vol. 46, pp. 1744-50, Oct. 1958.
49. J. E. Keith Smith, "A Decision-Theoretic Speaker Recognizer", 64th Meeting of Acoust. Soc. of Amer., 8 Nov. 1962.
50. G. P. Dineen, "Programming Pattern Recognition", Proc. of Western Joint Computer Conf., pp. 91-3, 1955.
51. J. Y. Lettin, H. R. Maturana, W. S. McCulloch, and W. S. Pitts, "What the Frog's Eye Tells the Frog's Brain", Proc. IRE, Vol. 47, pp. 1940-51, 1959.
52. "Bionics Symposium, Living Prototypes - The Key to New Technology", Wright Air Development Division Technical Report 60-600, December 1960.
53. H. Dudley and S. Balashuk, "Automatic Recognition of Phonetic Patterns in Speech; Phonetic Vocoder", Bell Tele. System Monograph p. 3172, 1958.
54. F. Rosenblatt, "Perception Simulation Experiments", Proc. IRE, pp. 301-309, March 1960.
55. A. Glovazky, "Determination of Redundancies in a Set of Patterns", IRE Trans., Vol. 1T2, pp. 151-3, Dec. 1956.
56. W. W. Peterson, "Error-Correcting Codes", John Wiley and Sons, New York, 1961.
57. P. W. Cooper, "Statistical Pattern Recognition with Quadratic Forms", Melpar Technical Note 62/4, Watertown Mass., June 1962.
58. A. Lubin, "Linear and Non-linear Discriminating Functions", Brit. Journ. of Psychology (Stat. Sect.), Vol. 3, Part 2, pp. 90-103, 1950.
59. A. Wald, "Statistical Decision Functions", John Wiley and Sons, New York, 1950.
60. P. W. Cooper, "The Hypersphere in Pattern Recognition", Melpar Tech. Note 62/1, Watertown Mass., Feb. 1962.
61. N. Wiener, "Extrapolation, Interpolation, and Smoothing Stationary Time Series", John Wiley and Sons, New York, 1949.
62. H. W. Bode and C. E. Shannon, "A Simplified Derivation of Linear Least-Square Smoothing and Prediction", Proc. IRE, Vol. 38, pp. 417, 1950.
63. C. V. Jakowatz, "Adaptive Waveform Recognition", Fourth Intl. Sym. on Information Theory, London, Sept. 1960.

64. R. D. Turner, "Operations Research on Recognition - The Generalized Search Process", (AFCRL 945), G. E. Advanced Electronics Center, Ithaca, New York, Nov. 1961.
65. M. W. Evans, C. K. Mc Elwain, and F. Van Hoosen, "Machine Correction of Garbled English Text", Lincoln Lab. Report 54G-0022, 1 June 1960.
66. A. E. Laemmel, "Linear Statistical Classification", Lincoln Lab. Report 54G-0021, 16 May 1960..
67. C. Steinberg and J. McBride, "Pattern Recognition in the Electrocardiogram", Airborne Inst. Lab., Deer Park, New York, advertisement, March 1961.
68. M. Eden and M. Halle, "The Characterization of Cursive Writing", Fourth Int'l. Sym. on Info. Theory, London, Sept. 1960.
69. G. S. Sebestyn, "Decision-Making Processes in Pattern Recognition", Macmillan, New York, 1962.

Distribution List

List A

| <u>Organization</u> | <u>No. of Copies</u> | <u>Organization</u> | <u>No. of Copies</u> |
|--|----------------------|---|----------------------|
| AFMTC (AFMTC Tech Library-MU-135) Patrick AFB, Fla. | 1 | Defense Documentation Center (DDC) Arlington Hall Station Arlington 12, Va. | 10 |
| AUL Maxwell AFB, Ala. | | Office of Scientific Intelligence Central Intelligence Agency 2430 E. Street N. W. Washington 25, D.C. | 1 |
| OAR (RROS, Col. John R. Fowler) Tempo D. 4th and Independence Avenue Washington 25, D.C. | 1 | Scientific and Technical Information Facility Attn: NASA Representative (S-AK/ DL) P.O. Box 5700 Bethesda, Maryland | 1 |
| AFOSR, OAR (SRYP) Tempo D 4 th and Independence Ave. Washington 25, D.C. | 1 | Director Langley Research Center National Aeronautics and Space Administration Langley Field, Va. | 1 |
| ASD (ASAPRD-Dist) Wright-Patterson AFB, Ohio | 1 | AF Missle Development Center (MDGRT) Holloman AFB, New Mexico | 1 |
| RADC (RAALD) Griffiss AFB, New York Attn: Documents Library | 1 | Hq. OAR (RROSP, Maj. Richard W. 1 Nelson) Washington 25, D.C. | 1 |
| Commanding General USASRDL Ft. Monmouth, N.J. Attn: Tech. Doc. Ctr. SIGRA/ SL-ADT | 1 | Department of the Army Office of the Chief Signal Officer Washington 25, D.C. Atten: SIGRD-4a-2 | 1 |
| Commanding Officer Attn: ORDTL-012 Diamond Ordnance Fuze Laboratories Washington 25, D.C. | 1 | Chief, Bureau of Naval Weapons Department of the Navy Wash. 25, D.C. Attn: DLI-31 | 2 |
| Redstone Scientific Information Center U.S. Army Missile Command Redstone Arsenal, Alabama | 1 | Director (Code 2027) U.S. Naval Research Laboratory Wash. 25, D.C. | 2 |
| | | Director, USAF Project RAND The Rand Corporation 1700 Main Street Santa Monica, Calif. THRU: AF Liaison Office | 1 |

Distribution List (continued)

| <u>Organization</u> | <u>No. of Copies</u> | <u>Organization</u> | <u>No. of Copies</u> |
|---|----------------------|--|----------------------|
| Director National Security Agency Fort George Meade, Maryland Attn: K-14 | 1 | Institute of the Aerospace Sciences, Inc. 2 E. 64th Street New York 21, N.Y. Attn: Librarian | 1 |
| Technical Information Office European Office, Aerospace Research Shell Building 47 Cantersteen Brussels, Belgium | 1 | Office of Naval Research Branch Office, London Navy 100, Box 39 F.P.O., N.Y., N.Y. | 2 |
| Aero Res. Lab. (OAR) AROL Lib. AFL 2292 Bldg. 450 Wright-Patterson, AFB, Ohio | 1 | Massachusetts Institute Of Technology Research Laboratory Building 26, Rm. 327 Cambridge 39, Mass. Attn: John H. Hewitt | 1 |
| U.S. Army Aviation Human Research Unit U.S. Continental Army Command P.O. Box 428, Fort Rucker, Alabama Attn: Maj. Arne H. Eliasson | 1 | Alderman Library University of Virginia Charlottesville, Va. | 1 |
| Library Boulder Laboratories National Bureau of Standards Boulder, Colorado | 2 | Defence Research Member Canadian Joint Staff 2450 Massachusetts Ave. N.W. Wash 8, D.C. ("Technical/scientific reports will be released for military purposes only and any propriety rights which may be involved are protected by U.S./ Canadian Government agreements.) | 3 |

List A-S

| | | | |
|--|---|--|---|
| WADD (WCLN) Wright-Patterson AFB, Ohio | 1 | Hq. USAF (AFDRD) Wash 25, D.C. | 2 |
| WADD (WCLR) Wright-Patterson AFB, Ohio | 1 | RADC (RCW) Griffiss AFB, N.Y. | 1 |
| RADC (RCU) Griffiss AFB, N.Y. | 1 | RADC (RCSGPR, Milton V. Ratynski) Griffiss AFB, N.Y. | 1 |
| WADD (WCLJ) Wright-Patterson AFB, Ohio | 1 | ESD (ESRHC) Attn: Dr. Joseph M. Doughty | 3 |
| RADC (RCE) Griffiss AFB, N.Y. | 1 | L.G. Hanscom Field Bedford, Mass. | |
| AFOSR (Dr. Harold Wooster) Director Research Information Office Wash 25, D.C. | 1 | Motorola, Inc. 8201 E. McDowell Rd. Phoenix, Arizona Attn: Dr. Thomas E. Tice | 1 |

Distribution List (continued)

| <u>Organization</u> | <u>No. of Copies</u> | <u>Organization</u> | <u>No. of Copies</u> |
|--|----------------------|--|----------------------|
| Bolt, Beranek and Newman 50 Moulton Street Cambridge 38, Mass. Attn: Dr. J. C. R. Licklider | 1 | University of Michigan Office of Research Administration Radiation Laboratory 912 N. Main Street Ann Arbor, Michigan Attn: Mr. Ralph Hiatt | 1 |
| General Electric Company Technical Military Planning Operation 735 State Street Santa Barbara, Calif. Attn: Mr. G. P. Mandanis | 1 | Librarian Departments of Physics University of Illinois Urbana, Illinois | 1 |
| The Mitre Corporation P. O. Box 208 Lexington 73, Mass Attn: Mr. R. R. Shorey, D-17 | 1 | Ohio State University 1314 Kinnear Road Columbus 8, Ohio Attn: Laboratory of Aviation Psychology | 1 |
| Planning Research Corp 10970 LeConte Avenue Los Angeles 24, Calif Attn: Mr. Irving Garfunkel | 1 | Northeastern Univ. 360 Huntington Avenue Boston, Mass. Attn: Prof. L. O. Dolansky | 1 |
| Cornell Aeronautical Lab., Inc. 4455 Genesee Street Buffalo 21, N.Y. Attn: Mr. E. W. Roth | 1 | Massachusetts Institute of Technology Dept. of Mechanical Engineering Cambridge 39, Mass. Attn: Prof. T. B. Sheridan Dynamic Analysis and Control Lab. | 1 |
| Office of Naval Research Department of the Navy Washington 25, D.C. Attn: Code 427; Electronics Branch | 1 | Northeastern Univ. 360 Huntington Avenue Boston 15, Mass. Attn: Prof. R. H. Moody Library | 1 |
| Commanding Officer and Director U.S. Navy Electronics Laboratory (Library) San Diego 52, Calif | 1 | Massachusetts Institute of Technology Operations Research Center Room 6-218 Cambridge 39, Mass | 1 |
| Commanding Officer U.S. Naval Ordnance Laboratory Corona, Calif. Attn: Documents Librarian | 1 | Case Institute of Technology University Circle Cleveland 6, Ohio Attn: B. V. Dean | 1 |
| Office of Naval Research Department of the Navy Washington 25, D.C. Attn: Code 418 | 1 | AFCRL, Office of Aerospace Research (CRUI) L.G. Hanscom Field, Bedford, Mass. | 10 |
| Massachusetts Institute of Technology P. O. Box 73 Lincoln Laboratory Lexington 73, Mass. Attn: M. A. Granese, Librarian | 1 | | |

UNCLASSIFIED

UNCLASSIFIED